Refine Search

Search Results -

Term	Documents
PHRASE	33677
PHRASES	11507
FREQUENCY	606566
FREQUENCIES	203125
FREQUENCYS	34
(19 AND (PHRASE NEAR FREQUENCY)).USPT.	0
(L19 AND (PHRASE NEAR FREQUENCY)).USPT.	0

Database:
US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Search: L25

Refine Search

Recall Text | Clear | Interrupt

Search History

DATE: Saturday, November 27, 2004 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
	DB=USPT; PLUR=YES; OP=OR		
L25	l19 and (phrase near frequency)	0	L25
L24	l21 and (phrase near frequency)	0	L24
L23	l22 and (phrase near frequency)	0	L23
L22	l21 and normaliz$	1	L22
L21	l19 and unstructur$	4	L21
L20	L19 and (normaliz$ near matrix)	0	L20
L19	L18 and taxonomy	9	L19
L18	L17 and extract$	113	L18

L17	l1 and "co-occurrence"	136	L17
L16	L1 and "co-occurrency"	0	L16
L15	L14 and "co-occurrency"	0	L15
L14	L13 and phrase	931	L14
L13	L1 and extract$	4861	L13
L12	l11 and taxonomy	0	L12
L11	l9 and "co-occurrence"	1	L11
L10	l9 and collection	0	L10
L9	l7 and (phrase near frequency)	2	L9
L8	l7 and (frequence near phrase)	0	L8
L7	(phrase near $$occurence)	2	L7
L6	L2 and (phrase near $$occurence)	0	L6
L5	(phase near $$occurence)	14	L5
L4	L2 and (phase near $$occurence)	0	L4
L3	L2 and (phase near frequency)	0	L3
L2	L1 and (document near retriev$)	1143	L2
L1	707/$.ccls.	13922	L1

END OF SEARCH HISTORY

L9: Entry 1 of 2 File: USPT Aug 8, 1995

DOCUMENT-IDENTIFIER: US 5440481 A
TITLE: System and method for database tomography

Detailed Description Paragraph Table (3):
TABLE 3 _____ 150(C.sub.j) REMOTE SENSING PTA - CLOSELY RELATED PHRASES C.sub.ij C.sub.i _____ 139 2764 DATA 022 0036 THERMAL INFRARED 120 0879 REMOTE 056 0323 ICE 079 2593 LITERATURE 070 0522 SATELLITE 041 0228 OCEANOGRAPHIC 012 0020 ATMOSPHERIC CORRECTIONS 065 2287 UNTIED 109 1707 SPACE 012 0024 AEROSOL OPTICAL 012 0025 IMAGING SYSTEMS 006 0007 MICROWAVE SENSORS 062 2239 UNITED STATES 074 1072 RADAR 012 0037 VEGETATION _____ CODE: C.sub.ij IS COOCCURRENCE FREQUENCY, OR NUMBER OF TIMES PHRASE APPEARS WITHIN +/- 50 WORDS OF PTA IN TOTAL TEXT; C.sub.i IS ABSOLUTE OCCURRENCE FREQUENCY OF PHRASE; C.sub.j IS ABSOLUTE OCCURRENCE FREQUENCY OF PTA.

Detailed Description Paragraph Table (4):
TABLE 4 _____ 150(C.sub.j) REMOTE SENSING PTA - CLOSELY RELATED PHRASES I.sub.i E.sub.ij C.sub.ij C.sub.i C.sub.ij C.sub.ij 2/C.sub.i C.sub.j PTA MEMBER _____ 022 0036 0.611 0.0359 THERMAL INFRARED 056 0323 0.173 0.0259 ICE 070 0522 0.134 0.0250 SATELLITE 041 0228 0.180 0.0197 OCEANOGRAPHIC 012 0020 0.600 0.0192 ATMOSPHERIC CORRECTIONS 109 1707 0.064 0.0186 SPACE 012 0024 0.500 0.0160 AEROSOL OPTICAL 006 0007 0.857 0.0137 MICROWAVE SENSORS 074 1072 0.069 0.0136 RADAR 012 0037 0.324 0.0104 VEGETATION _____ CODE: C.sub.ij IS COOCCURRENCE FREQUENCY, OR NUMBER OR TIMES PHRASE APPEARS WITHIN +/- 50 WORDS OF PTA IN TOTAL TEXT; C.sub.i IS ABSOLUTE OCCURRENCE FREQUENCY OF PHRASE; C.sub.j IS ABSOLUTE OCCURRENCE FREQUENCY OF PTA PHRASE; I.sub.i, THE INCLUSION INDEX BASED ON PHRASE, IS RATIO OF C.sub.ij TO C.sub.i ; AND E.sub.ij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON PHRASE I.sub.i (C.sub.ij /C.sub.i) AND INCLUSION INDEX BASED ON PTA I.sub.j (C.sub.ij /C.sub.j).

CLAIMS:

2. The system of claim 1 wherein said means for identifying pervasive theme areas in said database, comprises:

a means for counting frequency of occurrence of said n* word phrases;

a means for creating a list of all n* word phrases and the frequency of occurrence for each of said n* word phrases;

a means for sorting said list of n* word phrases by frequency of occurrence:

a means for defining pervasive theme areas from said list of sorted phrases; and

a means for selecting the number of said n* word phrases to be used as pervasive theme areas.

10. The computer implemented method of full-text database searching, comprising the steps of:

a. assembling information into a full-text database by scanning documents and storing digitized results in said computer;

b. eliminating trivial phrases from said databases by comparing a user-input list of such phrases with the entire contents of said database and deleting matches with said list;

c. using the definition of phrase as m*word=phrase where m=1,2,3, . . . n and where each word phrase for m=2,3 . . . n is composed of adjacent words, said word phrase for m=1 being a single word phrase, for m=2 an adjacent double word phrase, and for m=3 an adjacent triple word phrase, . . . and for m=n an adjacent nth word phrase, creating a list of all single word phrases, a list of all adjacent double word phrases, a list of all adjacent triple word phrases, . . . , and a list of all adjacent nth word phrases and their frequencies of occurrence in the database;

d. sorting each list of said phrases by their frequency of occurrence in said list;

e. identifying pervasive theme areas in the information in said database;

f. defining pervasive theme areas from said sorted list of phrases as the most frequently occurring phrases of high user-interest.

g. identifying phrases in said database that are related to said pervasive theme areas;

h. quantifying strength of relationships between said identified phrases and said pervasive theme areas;

i. identifying pervasive theme areas which are closely related;

j. displaying relationships among related pervasive theme areas and pervasive theme areas and related phrases; wherein the step of identifying phrases related to pervasive theme area further comprises the steps of

k. extracting phrases for each pervasive theme area (PTA) from the full-text database which occur within a user-identified range of interest plus or minus a range of words of the PTA; and

l. listing the extracted phrases and their frequency of occurrence in the database for each PTA.

13. The method of claim 12 wherein said step of identifying pervasive theme area further comprises the steps:

creating a list of all phrases sorted and ordered in accordance with frequency of occurence of said phrases;

sorting said ordered phrases in accordance with a user pre-specified user-interest criteria;

defining pervasive theme areas from said user interest phrases; and

selecting the number of said phrases to be used as pervasive theme areas.

Previous Doc Next Doc Go to Doc#

First Hit Fwd Refs Previous Doc Next Doc Go to Doc#
End of Result Set

Generate Collection Print

L9: Entry 2 of 2 File: USPT May 10, 1988

DOCUMENT-IDENTIFIER: US 4744050 A
TITLE: Method for automatically registering frequently used phrases

Detailed Description Text (12):
The frequent phrase processing program resides in the text processing program 208. It selects the phrases having high frequencies of occurence from the phrase table 16 shown in FIG. 1 in the text pocessing program 208, determines macro codes therefor and registers them in the macro table 13 shown in FIG. 1. It reads out the phrase corresponding to the keyed-in macro code and supplies it to the text processing program 208. In response to an instruction by a user, it sends the content of the macro table 13 to the terminal input/output control circuit 210 to display it on the screen 203.

Detailed Description Text (25):
The automatic registration of the frequent phrases by the macro table updating program 6 is now explained. FIG. 6 shows a flow chart of this processing. The processing may be carried out at any time although it is effective to carry out when the content of the phrase table 16 is changed. In the embodiment of FIG. 6, it is started when the number of times of phrase conversion exceeds a predetemined number in a step 81a. Thus, the registration is made at a constant time interval. Alternatively, the phrases having frequencies of occurrence higher than a predetermined number may be selected. In this case, they are registered at each time. The frequency field 37 and the flag 38 of the phrase table 16 are checked to search the phrase in the character string which has the flag value "0" and the highest frequency (step 81b). If the value of flag 38 is "0", it indicates that the phrase has not yet been registered in the macro table 13, and if it is "1", it indicates that it has been registered. Accordingly, the phrase having the highest frequency is selected from the unregistered phrases.

Detailed Description Text (28):
In the key mode, one character entered after the single phrase code designation key is processed in the essentially same manner as above. In the above embodiment, the registration processing is terminated if the table has no vacant area. Alternatively, one of the registered phrases having a lower frequency than that of the phrase to be registered may be deleted to make a vacant area. Only those phrase having lengths longer than a predetermined length may be automatically registered.

Previous Doc Next Doc Go to Doc#

L11: Entry 1 of 1 File: USPT Aug 8, 1995

DOCUMENT-IDENTIFIER: US 5440481 A
TITLE: System and method for database tomography

Brief Summary Text (6):
Modern quantitative techniques utilize computer technology extensively, usually supplemented by network analytic approaches, and attempt to integrate disparate fields of information. One class of techniques exploits the use of co-occurrence phenomena. In co-occurrence analysis, phenomena that occur together frequently in some domain are assumed to be related, and the strength of that relationship is assumed to be related to the co-occurrence frequency. Networks of these co-occurring phenomena are constructed, and then maps of evolving topic fields are generated using the link-node values of the networks. Using these maps of structure and evolution, the information analyst can develop a deeper understanding of the interrelationships among the different information fields and the impacts of external intervention, and can recommend new directions for more desirable information portfolios.

Brief Summary Text (7):
One approach to co-occurrence analysis is co-word analysis. The origins of Co-word phenomena can be traced back to the pioneering work in: 1) lexicography to account for co-occurrence knowledge, and 2) linguistics to describe how affinity of two language units correlates with their appearance in the language.

Brief Summary Text (8):
In early co-word studies, words were classified on the basis of their co-occurrence with other words as well as their meanings. It was, however, observed that the reasons for two words co-occurring in the same context are not always relevant to a general linguistic description of a given language. The well-formedness of sentences to their lexical levels; i.e., how sensitive the meaning of a sentence is to substitution for one member of co-occurrence pair has been studied. A recent study included collocations as part of a linguistic model, whose goal was to relate any given meaning and all the texts that express it. Information retrieval research has focused on designing more efficient indexing tools using pairwise lexical affinities instead of keywords. Methods have been developed for locating interesting collocational expressions in a large body of text. These methods were based principally on the distribution of types and tokens in the body of text and on the analysis of the statistical patterns of neighboring words.

Brief Summary Text (9):
In the mid-1970s, a study was performed to examine relationships among themes in a novel using co-occurrence phenomena. An important term in the book was chosen, and a dictionary was constructed of all words in the book occurring in the same sentences as that word. A co-occurrence matrix which contained the co-occurrences among these related terms was constructed, and analyzed to eventually show the relations among all the associated terms in the mini-dictionary as they occurred in the original text. While the dictionary was restricted to single words, and the co-occurrence domain was restricted to sentences, the methodology did represent a major step forward in extracting word relations from text by their co-occurrences.

Brief Summary Text (10):
A recent update of this method employed frequency of co-occurrence to extract relatedness information from text. The study looked at co-occurrence using the sense-definition as the

textual unit (entire definition of a sense of a word). The database used was the Longman Dictionary of Contemporary English (LDOCE) rather than free text. The method used single word frequencies only, and resulted in construction of networks of related words. It was concluded that co-occurrences of words in the LDOCE-controlled vocabulary in the definitions in LDOCE appeared to provide some useful information about the meanings of those words. Co-occurrences frequency correlated significantly with human judgements of relatedness, and the relatedness functions on co-occurrences yielded even higher correlations.

Detailed Description Paragraph Table (3):
TABLE 3 _____ 150(C.sub.j) REMOTE SENSING PTA - CLOSELY RELATED PHRASES C.sub.ij C.sub.i _____ 139 2764 DATA 022 0036 THERMAL INFRARED 120 0879 REMOTE 056 0323 ICE 079 2593 LITERATURE 070 0522 SATELLITE 041 0228 OCEANOGRAPHIC 012 0020 ATMOSPHERIC CORRECTIONS 065 2287 UNTIED 109 1707 SPACE 012 0024 AEROSOL OPTICAL 012 0025 IMAGING SYSTEMS 006 0007 MICROWAVE SENSORS 062 2239 UNITED STATES 074 1072 RADAR 012 0037 VEGETATION _____ CODE: C.sub.ij IS COOCCURRENCE FREQUENCY, OR NUMBER OF TIMES PHRASE APPEARS WITHIN +/- 50 WORDS OF PTA IN TOTAL TEXT; C.sub.i IS ABSOLUTE OCCURRENCE FREQUENCY OF PHRASE; C.sub.j IS ABSOLUTE OCCURRENCE FREQUENCY OF PTA.

Detailed Description Paragraph Table (4):
TABLE 4 _____ 150(C.sub.j) REMOTE SENSING PTA - CLOSELY RELATED PHRASES I.sub.i E.sub.ij C.sub.ij C.sub.i C.sub.ij C.sub.ij 2/C.sub.i C.sub.j PTA MEMBER _____ 022 0036 0.611 0.0359 THERMAL INFRARED 056 0323 0.173 0.0259 ICE 070 0522 0.134 0.0250 SATELLITE 041 0228 0.180 0.0197 OCEANOGRAPHIC 012 0020 0.600 0.0192 ATMOSPHERIC CORRECTIONS 109 1707 0.064 0.0186 SPACE 012 0024 0.500 0.0160 AEROSOL OPTICAL 006 0007 0.857 0.0137 MICROWAVE SENSORS 074 1072 0.069 0.0136 RADAR 012 0037 0.324 0.0104 VEGETATION _____ CODE: C.sub.ij IS COOCCURRENCE FREQUENCY, OR NUMBER OR TIMES PHRASE APPEARS WITHIN +/- 50 WORDS OF PTA IN TOTAL TEXT; C.sub.i IS ABSOLUTE OCCURRENCE FREQUENCY OF PHRASE; C.sub.j IS ABSOLUTE OCCURRENCE FREQUENCY OF PTA PHRASE; I.sub.i, THE INCLUSION INDEX BASED ON PHRASE, IS RATIO OF C.sub.ij TO C.sub.i ; AND E.sub.ij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON PHRASE I.sub.i (C.sub.ij /C.sub.i) AND INCLUSION INDEX BASED ON PTA I.sub.j (C.sub.ij /C.sub.j).

CLAIMS:

2. The system of claim 1 wherein said means for identifying pervasive theme areas in said database, comprises:

a means for counting frequency of occurrence of said n* word phrases;

a means for creating a list of all n* word phrases and the frequency of occurrence for each of said n* word phrases;

a means for sorting said list of n* word phrases by frequency of occurrence:

a means for defining pervasive theme areas from said list of sorted phrases; and

a means for selecting the number of said n* word phrases to be used as pervasive theme areas.

10. The computer implemented method of full-text database searching, comprising the steps of:

a. assembling information into a full-text database by scanning documents and storing digitized results in said computer;

b. eliminating trivial phrases from said databases by comparing a user-input list of such phrases with the entire contents of said database and deleting matches with said list;

c. using the definition of phrase as m*word=phrase where m=1,2,3, . . . n and where each word phrase for m=2,3 . . . n is composed of adjacent words, said word phrase for m=1 being a single word phrase, for m=2 an adjacent double word phrase, and for m=3 an adjacent triple word phrase, . . . and for m=n an adjacent nth word phrase, creating a list of all single word

phrases, a list of all adjacent double word phrases, a list of all adjacent triple word phrases, . . . , and a list of all adjacent nth word phrases and their frequencies of occurrence in the database;

d. sorting each list of said phrases by their frequency of occurrence in said list;

e. identifying pervasive theme areas in the information in said database;

f. defining pervasive theme areas from said sorted list of phrases as the most frequently occurring phrases of high user-interest.

g. identifying phrases in said database that are related to said pervasive theme areas;

h. quantifying strength of relationships between said identified phrases and said pervasive theme areas;

i. identifying pervasive theme areas which are closely related;

j. displaying relationships among related pervasive theme areas and pervasive theme areas and related phrases; wherein the step of identifying phrases related to pervasive theme area further comprises the steps of

k. extracting phrases for each pervasive theme area (PTA) from the full-text database which occur within a user-identified range of interest plus or minus a range of words of the PTA; and

l. listing the extracted phrases and their frequency of occurrence in the database for each PTA.

13. The method of claim 12 wherein said step of identifying pervasive theme area further comprises the steps:

creating a list of all phrases sorted and ordered in accordance with frequency of occurence of said phrases;

sorting said ordered phrases in accordance with a user pre-specified user-interest criteria;

defining pervasive theme areas from said user interest phrases; and

selecting the number of said phrases to be used as pervasive theme areas.

Previous Doc Next Doc Go to Doc#

Hit List

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Search Results - Record(s) 1 through 9 of 9 returned.

☐ 1. Document ID: US 6735592 B1

L19: Entry 1 of 9 File: USPT May 11, 2004

US-PAT-NO: 6735592
DOCUMENT-IDENTIFIER: US 6735592 B1

TITLE: System, method, and computer program product for a network-based content exchange system

DATE-ISSUED: May 11, 2004

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Neumann; Seth	Mountain View	CA		
Obsitnik; Steve	San Francisco	CA		
Whittemore; Greg	San Jose	CA		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Discern Communications	Menlo Park	CA			02

APPL-NO: 09/ 713520 [PALM]
DATE FILED: November 16, 2000

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/101; 707/102, 707/3, 709/206
US-CL-CURRENT: 707/101; 707/102, 707/3, 709/206

FIELD-OF-SEARCH: 707/101, 707/522, 707/3, 707/4, 707/5, 707/9, 707/10, 709/218, 709/104, 709/206, 709/102, 705/44

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
5774859	June 1998	Houser et al.	704/275
5774860	June 1998	Bayya et al.	704/275
5842163	November 1998	Weintraub	704/240
5873062	February 1999	Hansen et al.	704/254
5901287	May 1999	Bull et al.	395/200.48
5903882	May 1999	Asay et al.	705/44

5933822	August 1999	Braden-Harder et al.	707/5
5960399	September 1999	Barclay et al.	704/270
5991721	November 1999	Asano et al.	704/257
5995943	November 1999	Bull et al.	705/14
5996007	November 1999	Klug et al.	709/218
6052439	April 2000	Gerszberg et al.	379/88.01
6073102	June 2000	Block	704/275
6078924	June 2000	Ainsbury et al.	707/101
6081774	June 2000	de Hita et al.	704/9
6094649	July 2000	Bowen et al.	707/3
6101468	August 2000	Gould et al.	704/251
6105023	August 2000	Callan	707/5
6199082	March 2001	Ferrel et al.	707/522
6430602	August 2002	Kay et al.	707/10

ART-UNIT: 2175

PRIMARY-EXAMINER: Rones; Charles

ASSISTANT-EXAMINER: Mahmoudi; Hassan

ATTY-AGENT-FIRM: Moser, Patterson & Sheridan, LLP. Tong, Esq.; Kin-Wah

ABSTRACT:

A system, method and computer program product are provided for providing a content exchange system. A request is received from a user utilizing a local system. A determination is made as to whether the user request can be fulfilled from information stored by the local system. The request is fulfilled from a local data source if the request can be fulfilled locally. If the request cannot be fulfilled locally, the request is fulfilled at a network site. A content directory connected to the network site is examined for selecting one or more network data sites having content potentially satisfying the request. The request is sent to the data site (s). Content pertaining to the request is received from the data site(s) and sent to the user.

21 Claims, 13 Drawing figures

☐ 2. Document ID: US 6687696 B2

L19: Entry 2 of 9 File: USPT Feb 3, 2004

US-PAT-NO: 6687696
DOCUMENT-IDENTIFIER: US 6687696 B2

TITLE: System and method for personalized search, information filtering, and for generating recommendations utilizing statistical latent class models

DATE-ISSUED: February 3, 2004

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Hofmann; Thomas	Barrington	RI		
Puzicha; Jan Christian	Albany	CA		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Recommind Inc.	Berkeley	CA			02

APPL-NO: 09/ 915755 [PALM]
DATE FILED: July 26, 2001

PARENT-CASE:
This application claims the benefit of U.S. Provisional application No. 60/220,926, filed Jul. 26, 2000. Application Serial No. 60/220,926 is hereby incorporated by reference.

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/6; 707/4
US-CL-CURRENT: 707/6; 707/4

FIELD-OF-SEARCH: 707/1, 707/100, 707/101, 707/104.1, 707/500, 707/3, 707/4, 707/10, 707/200, 707/201, 707/6, 709/203, 709/217, 704/1, 704/9, 704/10, 703/22, 703/10, 705/26

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
5278980	January 1994	Pedersen et al.	707/4
5704017	December 1997	Heckerman et al.	
5724567	March 1998	Rose et al.	
5790426	August 1998	Robinson	
5790935	August 1998	Payton	
5867799	February 1999	Lang et al.	
5884282	March 1999	Robinson	
5918014	June 1999	Robinson	
5983214	November 1999	Lang et al.	
6006218	December 1999	Breese et al.	
6029141	February 2000	Bezos et al.	
6029195	February 2000	Herz	
6041311	March 2000	Chislenko et al.	
6049777	April 2000	Sheena et al.	
6064980	May 2000	Jacobi et al.	
6072942	June 2000	Stockwell et al.	
6078740	June 2000	DeTreville	
6138116	October 2000	Kitagawa et al.	
6493702	December 2002	Adar et al.	707/3
6510406	January 2003	Marchisio	704/9

OTHER PUBLICATIONS

T. Hofmann and J. Puzicha, Statistical Models for Co-occurrence DataTechnical Report 1625, MIT, 1998.
S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science,1990.
T. Hofmann, Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization, Advances in Neural Information Processing Systems 12, pp. 914-920, MIT Press, Jun. 2000.
Patrick Baudisch, Joining Collaborative And Content-Based Filtering, CHI '99 Workshop: Interacting with Recommender Systems, 1999.
S.T. Dumais, Latent Semantic Indexing (LSI), Proceedings of the Text Retrieval conference (TREC-3)), pp. 219-230, 1995.
F. Pereira, N. Tishby and L. Lee, Distributional Clustering of English Words, Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 183-190, 1993.
M. Evans, Z. Gilula and I. Guttman, Latent Class Analysis of Two-Way Contingency Tables by Bayesian Methods, Biometrika, V. 76, No. 3, pp. 557-563, 1989.
Z. Gilula, S. Haberman, Canonical Analysis of Contingency Tables of Maximum Likelihood, Journal of the American Statistical Association, V. 81, No. 395, pp. 780-788, 1986.
T. Hofmann, J. Puzicha and M. I. Jordan, Learning from Dyadic Data, Advances in Neural Information Processing Systems vol. 11, MIT Press. 1999.
K. Rose, E. Gurewitz, and G. Fox, A Deterministic Annealing Approaches Clustering, Pattern Recognition Letters 11, pp. 589-594, 1990.
D. Lee and S. Seung Learning The Parts Of Objects By Non-Negative Matrix Factorization Nature, vol. 401, pp. 788-791 1999.
D. Gildea and T. Hofmann, Topic-Based Language Models Using EM, Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH), 1999.
L. Saul and F. Pereira, Aggregate And Mixed-Order Markov Models For Statistical Language Processing, Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing, 1997.
A. Rao, D. Miller, K. Rose, and A. Gersho, Deterministically annealed mixture of experts models for statistical regression, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 3201-3204, IEEE Comput. Soc. Press, 1997.
L. H. Ungar and D. P. Foster, Clustering Methods For Collaborative Filtering, AAAI Workshop on Recommendation Systems, 1998.
L. H. Ungar and D. P. Foster, A Formal Statistical Approach To Collaborative Filtering, Proceedings of Conference on Automated Leading and Discovery (CONALD), 1998.
L. D. Baker and A. K. McCallum, Distributional Clustering Of Words For Text Classification, SIGIR, 1998.
J. S. Breese, D. Heckerman, and C. Kadie, Empirical Analysis Of Predictive Algorithms For Collaborative Filtering, Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998.
D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, Using Collaborative Filtering To Weave An Information Tapestry, Communications of the ACMV. 35, No. 12, pp. 61-70, 1992.
T. K. Landauer and S. T. Dumais, A Solution To Plato's Problem: The Latent Semantic Analysis Theory Of Acquisition, Induction, And Representation Of Knowledge, Psychological Review, V. 104, No. 2, pp. 211-240, 1997.
A. P. Dempster; N. M. Laird; and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal Royal Statistical Society, V. 39, pp. 1-38, 1977.

ART-UNIT: 2175

PRIMARY-EXAMINER: Rones; Charles

ASSISTANT-EXAMINER: Abel-Jalil; Naveen

ATTY-AGENT-FIRM: Hahn Loeser & Parks, LLP Minns; Michael H.

ABSTRACT:

The disclosed system implements a novel method for personalized filtering of information and

Hit List

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Search Results - Record(s) 1 through 2 of 2 returned.

☐ 1. Document ID: US 5440481 A

L9: Entry 1 of 2 File: USPT Aug 8, 1995

US-PAT-NO: 5440481
DOCUMENT-IDENTIFIER: US 5440481 A

TITLE: System and method for database tomography

DATE-ISSUED: August 8, 1995

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Kostoff; Ronald N.	Falls Church	VA		
Miles; David L.	Ridgecrest	CA		
Eberhart; Henry J.	Ridgecrest	CA		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
The United States of America as represented by the Secretary of the Navy	Washington	DC			06

APPL-NO: 07/ 967341 [PALM]
DATE FILED: October 28, 1992

INT-CL: [06] G06 F 17/20, G06 F 17/30

US-CL-ISSUED: 364/419.08; 364/419.07, 364/419.19
US-CL-CURRENT: 707/5

FIELD-OF-SEARCH: 364/419.08, 364/419.07, 364/419.13, 364/419.19, 395/600

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
4839853	June 1989	Deerwester et al.	364/DIG.2
4849898	July 1989	Adi	364/419.08
4942526	July 1990	Okajima et al.	364/419
4992972	February 1991	Brooks et al.	364/DIG.2
4994967	February 1991	Asakawa	364/419.08
5056021	October 1991	Ausborn	364/419

5070456	December 1991	Garneau et al.	364/418.08
5146405	September 1992	Church	364/419.08
5276616	January 1994	Koga et al.	364/419.08
5280573	January 1994	Kuga et al.	395/145
5295261	March 1994	Simonetti	395/600
5301109	April 1994	Landauer et al.	364/419.19
5311429	May 1994	Tominaga	364/419.01
5321833	June 1994	Chang et al.	395/600
5325298	June 1994	Gallant	364/419.19
5333313	July 1994	Heising	395/600

ART-UNIT: 231

PRIMARY-EXAMINER: Huntley; David M.

ASSISTANT-EXAMINER: Bodendorf; A.

ATTY-AGENT-FIRM: Sliwka; Melvin J. Forrest, Jr.; John L.

ABSTRACT:

A Process for analyzing full-text is provided for identifying often-repea, high user interest, word phrases in a database. Often-repeated, high user interest, word phrases are defined as pervasive theme areas (PTAs). The process also allows the relationship defined as connectivity among the various PTAs to be identified. In addition, phrases that are in proximity to the PTAs and which are strongly supportive of the PTAs are identified. Numerical indices, figure of merit, and user defined thresholds are used to quantify relations between PTAs and among PTAs and phrases.

14 Claims, 5 Drawing figures

☐ 2. Document ID: US 4744050 A

L9: Entry 2 of 2 File: USPT May 10, 1988

US-PAT-NO: 4744050
DOCUMENT-IDENTIFIER: US 4744050 A

TITLE: Method for automatically registering frequently used phrases

DATE-ISSUED: May 10, 1988

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Hirosawa; Toshio	Machida			JP
Itoh; Tutomu	Hachioji			JP
Uehara; Tetsuzou	Tokyo			JP
Kurihara; Junichi	Hachioji			JP

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Hitachi, Ltd.	Tokyo			JP	03

APPL-NO: 06/ 748907 [PALM]
DATE FILED: June 26, 1985

FOREIGN-APPL-PRIORITY-DATA:

COUNTRY	APPL-NO	APPL-DATE
JP	59-129950	June 26, 1984

INT-CL: [04] G06F 15/38

US-CL-ISSUED: 364/900; 364/419
US-CL-CURRENT: 704/4; 715/531

FIELD-OF-SEARCH: 364/2MSFile, 364/9MSFile, 364/419

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
4339806	July 1982	Yoshida	364/900
4438505	March 1984	Yanagiuchi et al.	364/900
4468756	August 1984	Chan	364/900
4481607	November 1984	Kobayashi et al.	364/900
4499554	February 1985	Kobayashi	364/900
4502128	February 1985	Okajima et al.	364/900
4544276	October 1985	Horodeck	400/110
4559598	December 1985	Goldwasser et al.	364/419
4567573	January 1986	Hashimoto et al.	364/900
4630235	December 1986	Hashimoto et al.	364/900
4641264	February 1987	Nitta et al.	364/900

ART-UNIT: 237

PRIMARY-EXAMINER: Shaw; Gareth D.

ASSISTANT-EXAMINER: Napiorkowski; Maria

ATTY-AGENT-FIRM: Antonelli, Terry & Wands

ABSTRACT:

Phrases in an input character string are registered in a table (phrase table), frequencies of occurrence of those phrases are counted and registered in the table, the phrases having high frequencies of occurrence are selected and macro codes therefor are determined, and those macro codes and the corresponding phrases are paired and registered in the macro table. A content of the macro table thus prepared may be displayed on a display screen. Thereafter, a user can enter a desired phrase by keying the corresponding macro code.

9 Claims, 12 Drawing figures

Full | Title | Citation | Front | Review | Classification | Date | Reference | Claims | KWIC | Draw. Desc | Image

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Term	Documents
PHRASE	33677
PHRASES	11507
FREQUENCY	606566
FREQUENCIES	203125
FREQUENCYS	34
(7 AND (PHRASE NEAR FREQUENCY)).USPT.	2
(L7 AND (PHRASE NEAR FREQUENCY)).USPT.	2

Display Format: FRO Change Format

Previous Page Next Page Go to Doc#

L22: Entry 1 of 1 File: USPT Feb 3, 2004

DOCUMENT-IDENTIFIER: US 6687696 B2
TITLE: System and method for personalized search, information filtering, and for generating recommendations utilizing statistical latent class models

Brief Summary Text (10):
The disclosed system provides a method for the personalized filtering of information and the automated generation of user-specific recommendations. The system goes through 3 phases: 1) Information Gathering, 2) System Learning, and 3) Information Retrieval. The disclosed system is concerned primarily with the final two phases (System Learning and Information Retrieval). In the Information Gathering phase, information about the data to be retrieved (DOCUMENT DATA) and about the individual users (USER DATA) is collected. The USER DATA can be gathered explicitly through questionnaires, etc. or can be implied though observing user behavior such as Internet history logs, demographic information, or any other relevant sources of information. The DOCUMENT DATA can be gathered though a variety of methods including Internet crawling, topic taxonomies or any other relevant source of information. Once the Information Gathering phase is completed, the System Learning phase is initiated. The system employs a statistical algorithm that uses available USER DATA and DOCUMENT DATA to create a statistical latent class model (MODEL), also known as Probabilistic Latent Semantic Analysis (PLSA). The system learns one or more MODELS based on the USER DATA, DOCUMENT DATA, and the available database containing data obtained from other users. Within the MODEL, probabilities for words extracted from training data are calculated and stored in at least one matrix. An extended inverted index may also be generated and stored along with the MODEL in order to facilitate more efficient information gathering. The MODEL may be used in other applications such as the unsupervised learning of topic hierarchies and for other data mining functions such as identifying user communities. Various parts of the Information Gathering phase and the System Learning phase are repeated from time to time in order to further refine or update the MODEL. This refined or updated model will result in even higher levels of accuracy in processing the user's query. The final phase is the Information Retrieval phase. The user may enter a query. Once the query is entered into the system, the MODEL is utilized in calculating probabilities for every word in a document based upon at least 1) the user query, or 2) words associated with the users query in the MODEL, or 3) document information. All of the probabilities for a given document are added together yielding a total relevance "score" after which related documents are compared using this relevance score. The results are returned in descending order of relevance organized into at least one result list. Through the use of the MODEL, the system provides two benefits to the user: 1) the search results are personalized as each MODEL may be created in part using USER DATA, and 2) results for new users are somewhat personalized from the initial use through collaborative filtering based upon USER DATA for other system users.

Detailed Description Text (4):
The information available in typical information filtering applications is highly diverse. Thus, we are first concerned with abstracting away from this diversity in order to identify a few fundamental types of observation data. Co-occurrence data refers to a domain with two or more finite sets of objects in which observations are made for joint occurrences of objects, i.e., typically consist of tuples with one element from either set. This includes event dyadic data, histogram data, and single stimulus preference data as special cases. Co-occurrence data arises naturally in many applications ranging from computational linguistics and information retrieval to preference analysis and computer vision.

Detailed Description Text (5):
In online information filtering applications, we find three fundamental variables, namely objects (documents or products) o.epsilon.O, users or customers u.epsilon.U, and a vocabulary w.epsilon.W of terms and descriptors. Here O, U and W are discrete spaces (i.e. the set of all objects, the set of all users and the underlying vocabulary) so observations can be modeled as co-occurrences of these basic variables, e.g. user queries as (u;w.sub.1, . . . , w.sub.n), object description (o;w.sub.1, . . . , w.sub.n), buying events (u, o) etc.

Detailed Description Text (6):
More formally, our starting point is an observation sequence S=(x.sub.I.sub..sup.n .sup.n), 1.ltoreq.n.ltoreq.N, which is a realization of an underlying sequence of random variables (X.sub.I.sub..sup.n .sup.n) 1.ltoreq.n.ltoreq.N. Superscript indices are used to number observations. Capital letters without superscripts X are used to refer to generic instances of random variables with values in O, U or W, and X.sub.I refer to a generic co-occurrence where I is a multi-index for addressing the different co-occurring variables involved. In the modeling approach, it is assumed that the choice of a specific type of co-occurrence I.sup.n for the n-th observation is predetermined and is not part of the statistical model.

Detailed Description Text (7):
In this fashion, information filtering can be viewed in a statistical setting as completion of co-occurrences based on partial information, in essence the prediction of a response variable given a set of predictors. For example, a search engine functionality is modeled as predicting o given w.sub.1, . . . , w.sub.n or, in a statistical setting as computing the probability P (o.vertline.w.sub.1, . . . , w.sub.n) of observing o given the query terms. A recommender system is implemented by computing P(o.vertline.u) while P(o.vertline.u;w.sub.1, . . . , w.sub.n) implements a personalized search engine. Several other possible applications seem reasonable. FIG. 1 is a table showing possible queries in a combined content/collaborative system (taken in part from Patrick Baudisch, Joining collaborative and content-based filtering. CHI'99 Workshop: Interacting with Recommender Systems, 1999.) FIG. 1 provides an overview of possible modalities in a combined content/collaborative system where the rows relate to a query and the columns relate to a target. The middle row 2 of FIG. 1 represents actual recommender functionality, where users are interested in retrieving objects. The last row 4 is of special interest for marketing applications, where users are to be identified.

Detailed Description Text (8):
The key problem in the statistical analysis of co-occurrence data is data sparseness. While counts n(x.sub.I)=.vertline.x.sub.I.sub..sup.n .sup.n :x.sub.I.sub..sup.n .sup.n =x.sub.I.vertline. of the empirical frequency of an event x.sub.I capture all that can be possibly measured from the data, these sufficient statistics are subject to statistical fluctuations that for large underlying spaces and higher order co-occurrences become overwhelming, and therefore a direct estimation of joint occurrence probabilities becomes prohibitive.

Detailed Description Text (11):
Second, we then introduce the full, flat Probabilistic Latent Semantic Analysis model for generic multiway co-occurrence data that can be used, e.g. for joint collaborative and content filtering. While for the special case Probabilistic Latent Semantic Indexing several relationships to known proposals can be drawn, no competing approach is known for the full Probabilistic Latent Semantic Analysis method.

Detailed Description Text (14):
The starting point for Probabilistic Latent Semantic Indexing is a statistical model, which is also called (dyadic) aspect model. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable a.epsilon.A={a.sub.1, . . . , a.sub.K } with each observation. The modeling principle of latent variable models is the specification of a joint probability distribution for latent and observable variables. This unifies statistical modeling and structure detection: a probabilistic model of the observables is obtained by marginalization, while Bayes' rule induces posterior probabilities on the latent space of structures with respect to given observations. The latter provides a natural solution for topic extraction, word sense disambiguation and cataloging which correspond to different

values of the hidden variables. As a key advantage, mixture models provide both, a parsimonious yet flexible parameterization of probability distributions with good generalization performance on sparse data, as well as structural information about data-inherent grouping structure, which is discussed in detail below in the section entitled `The Cluster-Abstraction Model`. In the plain Probabilistic Latent Semantic Indexing model a joint probability model over OxW is defined by the mixture: ##EQU1##

Detailed Description Text (23):
In Latent Semantic Indexing, this is the L.sub.2 - or Frobenius norm, which corresponds to an implicit additive Gaussian noise assumption on (possibly transformed) counts. In contrast, Probabilistic Latent Semantic Indexing relies on the likelihood function of multinomial sampling and aims at an explicit maximization of the predictive power of the model. As is well known, this corresponds to a minimization of the cross entropy or Kullback-Leibler divergence between the empirical distribution and the model, which is very different from any type of squared deviation. On the modeling side this offers important advantages, for example, the mixture approximation P of the co-occurrence table is a well-defined probability distribution and factors have a clear probabilistic meaning. In contrast, Latent Semantic Indexing does not define a properly normalized probability distribution and may even contain negative entries. In addition, there is no obvious interpretation of the directions in the Latent Semantic Indexing latent space, while the directions in the Probabilistic Latent Semantic Indexing space are interpretable as multinomial word distributions.

Detailed Description Text (30):
The aspect model for multivariate co-occurrence data is built on the assumption that all co-occurrences in the sample S=(x.sub.I.sub..sup.n .sup.n) 1.ltoreq.n.ltoreq.N are independent and identically distributed and that random variables X.sub.i.sup.n and X.sub.j.sup.n i.noteq.j are conditionally independent given the respective latent class. The randomized data generation process can be described as follows: (i) Choose an aspect a with probability P(A=a) (or, in short notation, P(a)), and (ii) Select x.epsilon.{O,U,W} for all i.epsilon.I with probability P (x.sub.1.vertline.a).

Detailed Description Text (32):
By summing over all possible realizations of the latent variables and grouping identical co-occurrences together, one obtains the usual mixture probability distribution on the observables, ##EQU9##

Detailed Description Text (38):
After introducing appropriate Lagrange multipliers to ensure the correct normalization one obtains for the M-step formulae ##EQU13##

Detailed Description Text (51):
For prediction, we are interested in calculating probabilities P(x.sub.I.vertline.x.sub.J). Assuming we are interested in computing the probability of an object given a query and a user, P(o.vertline.w.sub.1, . . . , w.sub.n ;u). The first difficulty arises from the fact that we basically train models for co-occurrences with single w, so we assume conditional independence of keywords given object and user, ##EQU18##

Detailed Description Text (72):
When observing user behavior and preferences one may have richer observations than plain co-occurrences. Many use cases may also provide some additional preference value v with an observation. In this invention, we will treat the simplest case, where v.epsilon.{-1,+1} corresponds to either a negative or a positive example of preference, modeling events like "person u likes/dislikes object o".

Detailed Description Text (83):
FIGS. 9 through 13 illustrate an implementation of latent class models for personalized information filtering and recommendations. FIG. 9 shows the primary input streams into server platform 100. User profiles 111 are processed through a profiler module 110 to provide the user related information, such as transaction data, click stream data, download, demographic information, etc. Document information, such as text documents, web pages, emails, etc. comes

from a content repository 121 and is processed by a preprocessor and crawler module 120. Content repository 121 can be single proprietary database owned by the user. It can be any collection of data sources including any and all of the information available on the World Wide Web. The final primary input stream is expert annotations 131, including taxonomies, web catalogues, XML ontology, etc. and is processed by XML-parsing module 130. Here we assume that expert annotations of documents are stored as XML tags with the documents. Although this is common nowadays, other interfaces are of course possible.

Detailed Description Text (84):
FIG. 10 illustrates the main data processing modules for latent class models for personalized information filtering and recommendations. The concept extraction module 140 automatically extracts domain-specific concepts and topics for the documents provided by the preprocessor and crawler module 120. Preferably, this extraction includes statistically analyzing the data to learn the semantic associations between words within specific items in the acquired data. Also, preferably, probabilities for each learned semantic association are computed. The collaborative filtering module 142 analyzes the user profiles 111 provided by profiler 110 to learn about user interests long-term information needs. The collaborative filtering module performs 142 two functions: 1) it analyzes the current user's historical profile; and 2) it analyzes other users' historical profiles. Preferably both profile analyses are used in combination with the learned semantic associations and computed probabilities to provide improved predictions or recommendations lists. The categorization module 144 automatically annotates documents with appropriate categories. The data mining module 145 extracts taxonomies, topic maps and user communities. Depending upon the needs of the user one or more of these modules 140, 142, 144, 145 are used to implement latent class modeling for personalized information filtering and recommendations. All four modules are not necessarily used for each implementation.

Detailed Description Text (85):
FIG. 11 illustrates some of the preferred applications of the present invention. Functions included in the software provided by the server platform 100 are intelligent retrieval, categorization, filtering and recommendation. The intelligent retrieval of information incorporates user information from a user profile and from collaborative filtering into the search. From these functions, the present invention can provided personalized search results, automatic categorization of documents, email and text sorting and recommendations. The automatic categorization of documents categorizes the data into existing taxonomies and subject heading classification schemas. The email/text sorting can be used for intelligent information routing for customer relationship management (CRM) supply chain, distributed networking, storage, eMarketplaces, and web application server environments.

Detailed Description Text (86):
FIG. 12 illustrates one implementation of the present invention. A query 151 is input into server 100. The server 100 identifies matching documents 153 based on pure content analysis, it then connects the personalization engine 152 to access the user profile 111. Using the query 151 and the user profile 111, the server 100 uses the full probabilistic latent semantic analysis of user (community) data and content, 156 to produces an unstructured result list 155 with associated relevance scores. The taxonomy 154 is accessed to apply relevant concepts and categories to refine the unstructured result 155 into a structured search result 157. The structured search result can be further refined by including similar documents, refinement of the query by the user, etc.

Detailed Description Text (88):
Before processing the user's request, the server 100 analyzes the data collection and automatically extracts concepts, topics, and word contexts that are fully adapted to the specific data and the specific domain. The server 100 extracts concepts from the documents automatically, thus removing the need to manually tag documents. Most document collections are based on a very specific vocabulary, terminology or jargon. Contrary to standard information retrieval engines based on general dictionaries or thesauri, server 100 automatically creates indices and adapts and optimizes with respect to a specific document base. The server 100 architecture is divided into three main components: the learning module, the prediction module, and the data store. The learning module takes content (emails, documents, web-pages, data), adds it to the data store or document server 103 and makes this content accessible for

searching, categorizing, filtering, and delivery. The prediction module is used to perform searches against indexed documents. A search is entered using a web search page. The prediction module reduces the search to a set of documents that most clearly match the criteria of the search, and formats this set into a series of document lists, segmented by category. Because the system knows the different meanings of words, it can accommodate ambiguities by returning multiple result lists along with characterizing keywords. For example, the system can distinguish between Apple, the computer company; apple, the fruit; and Apple, the Beatles record company, and group the search results into three distinct lists. Users can then refine or rephrase queries.

Current US Original Classification (1):
707/6

Current US Cross Reference Classification (1):
707/4

Other Reference Publication (1):
T. Hofmann and J. Puzicha, Statistical Models for Co-occurrence DataTechnical Report 1625, MIT, 1998.

CLAIMS:

12. The method according to claim 1, further comprising: extracting hierarchical relationships between groups of data.

Previous Doc Next Doc Go to Doc#

Terms used **phrase frquency** and **phrase co occurrence** and **taxonomy** and **normalization**

Found **7,683** of **147,060**

Sort results by: relevance

Display results: expanded form

Save results to a Binder

Search Tips

☐ Open results in a new window

Try an Advanced Search

Try this search in The ACM Guide

Results 1 - 20 of 200 Result page: **1** 2 3 4 5 6 7 8 9 10 next

Best 200 shown

Relevance scale

1 Semantic indexing for a complete subject discipline

Yi-Ming Chung, Qin He, Kevin Powell, Bruce Schatz

August 1999 **Proceedings of the fourth ACM conference on Digital libraries**

Full text available: pdf(256.74 KB) Additional Information: full citation, references, index terms

Keywords: MEDLINE, MEDSPACE, concept space, interspace, medical informatics, scalable semantics, semantic indexing, semantic retrieval

2 Machine learning in automated text categorization

Fabrizio Sebastiani

March 2002 **ACM Computing Surveys (CSUR)**, Volume 34 Issue 1

Full text available: pdf(524.41 KB) Additional Information: full citation, abstract, references, citings, index terms

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. ...

Keywords: Machine learning, text categorization, text classification

3 Models of translational equivalence among words

I. Dan Melamed

June 2000 **Computational Linguistics**, Volume 26 Issue 2

Full text available: pdf(1.89 MB) Publisher Site Additional Information: full citation, abstract, references, citings

Parallel texts (bitexts) have properties that distinguish them from other kinds of parallel data. First, most words translate to only one other word. Second, bitext correspondence is typically only partial---many words in each text have no clear equivalent in the other text. This article presents methods for biasing statistical translation models to reflect these

properties. Evaluation with respect to independent human judgments has confirmed that translation models biased in this fashion are si ...

4 Generalizing case frames using a thesaurus and the MDL principle
Hang Li, Naoki Abe
June 1998 **Computational Linguistics**, Volume 24 Issue 2

Full text available: pdf(1.74 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

A new method for automatically acquiring case frame patterns from large corpora is proposed. In particular, the problem of generalizing values of a case frame slot for a verb is viewed as that of estimating a conditional probability distribution over a partition of words, and a new generalization method based on the Minimum Description Length (MDL) principle is proposed. In order to assist with efficiency, the proposed method makes use of an existing thesaurus and restricts its attention to thos ...

5 Learning to find answers to questions on the Web
Eugene Agichtein, Steve Lawrence, Luis Gravano
May 2004 **ACM Transactions on Internet Technology (TOIT)**, Volume 4 Issue 2

Full text available: pdf(4.49 MB) Additional Information: full citation, abstract, references, index terms

We introduce a method for learning to find documents on the Web that contain answers to a given natural language question. In our approach, questions are transformed into new queries aimed at maximizing the probability of retrieving answers from existing information retrieval systems. The method involves automatically learning phrase features for classifying questions into different types, automatically generating candidate query transformations from a training set of question/answer pairs, and ...

Keywords: Web search, information retrieval, meta-search, query expansion, question answering

6 Special issue of the lexicon: Tools and methods for computational lexicology
Roy J. Byrd, Nicoletta Calzolari, Martin S. Chodorow, Judith L. Klavans, Mary S. Neff, Omneya A. Rizk
July 1987 **Computational Linguistics**, Volume 13 Issue 3-4

Full text available: pdf(2.49 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

This paper presents a set of tools and methods for acquiring, manipulating, and analyzing machine-readable dictionaries. We give several detailed examples of the use of these tools and methods for particular analyses. A novel aspect of our work is that it allows the combined processing of multiple machine-readable dictionaries. Our examples describe analyses of data from Webster's Seventh Collegiate Dictionary, the Longman Dictionary of Contemporary English, the Collins bilingual dictionaries, t ...

7 Technique for automatically correcting words in text
Karen Kukich
December 1992 **ACM Computing Surveys (CSUR)**, Volume 24 Issue 4

Full text available: pdf(6.23 MB) Additional Information: full citation, abstract, references, citings, index terms, review

Research aimed at correcting words in text has focused on three progressively more difficult problems:(1) nonword error detection; (2) isolated-word error correction; and (3) context-dependent work correction. In response to the first problem, efficient pattern-matching and

n-gram analysis techniques have been developed for detecting strings that do not appear in a given word list. In response to the second problem, a variety of general and application-specific spelling cor ...

Keywords: n-gram analysis, Optical Character Recognition (OCR), context-dependent spelling correction, grammar checking, natural-language-processing models, neural net classifiers, spell checking, spelling error detection, spelling error patterns, statistical-language models, word recognition and correction

8 TextTiling: segmenting text into multi-paragraph subtopic passages
Marti A. Hearst
March 1997 **Computational Linguistics**, Volume 23 Issue 1

Full text available: pdf(2.46 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

TextTiling is a technique for subdividing texts into multi-paragraph units that represent passages, or subtopics. The discourse cues for identifying major subtopic shifts are patterns of lexical co-occurrence and distribution. The algorithm is fully implemented and is shown to produce segmentation that corresponds well to human judgments of the subtopic boundaries of 12 texts. Multi-paragraph subtopic segmentation should be useful for many text analysis tasks, including information retrieval and ...

9 Full Technical Papers: Learning implicit user interest hierarchy for context in personalization
Hyoung R. Kim, Philip K. Chan
January 2003 **Proceedings of the 8th international conference on Intelligent user interfaces**

Full text available: pdf(191.53 KB) Additional Information: full citation, abstract, references, index terms

To provide a more robust context for personalization, we desire to extract a continuum of general (long-term) to specific (short-term) interests of a user. Our proposed approach is to learn a user interest hierarchy (UIH) from a set of web pages visited by a user. We devise a divisive hierarchical clustering (DHC) algorithm to group words (topics) into a hierarchy where more general interests are represented by a larger set of words. Each web page can then be assigned to nodes in the hierarchy f ...

Keywords: clustering algorithm, concept clustering, user interest hierarchy, user profile

10 Towards language independent automated learning of text categorization models
Chidanand Apté, Fred Damerau, Sholom M. Weiss
August 1994 **Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(675.23 KB) Additional Information: full citation, references, citings, index terms, review

11 Term clustering of syntactic phrases
D. D. Lewis, W. B. Croft
December 1989 **Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(1.62 MB) Additional Information: full citation, abstract, references, citings, index terms

Term clustering and syntactic phrase formation are methods for transforming natural

language text. Both have had only mixed success as strategies for improving the quality of text representations for document retrieval. Since the strengths of these methods are complementary, we have explored combining them to produce superior representations. In this paper we discuss our implementation of a syntactic phrase generator, as well as our preliminary experiments with producing phrase clusters. Th ...

12 A survey of Web metrics

Devanshu Dhyani, Wee Keong Ng, Sourav S. Bhowmick

December 2002 **ACM Computing Surveys (CSUR)**, Volume 34 Issue 4

Full text available: pdf(289.28 KB) Additional Information: full citation, abstract, references, index terms

The unabated growth and increasing significance of the World Wide Web has resulted in a flurry of research activity to improve its capacity for serving information more effectively. But at the heart of these efforts lie implicit assumptions about "quality" and "usefulness" of Web resources and services. This observation points towards measurements and models that quantify various attributes of web sites. The science of measuring all aspects of information, especially its storage and retrieval or ...

Keywords: Information theoretic, PageRank, Web graph, Web metrics, Web page similarity, quality metrics

13 Session: On learning more appropriate Selectional Restrictions

Francesc Ribas

March 1995 **Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics**

Full text available: pdf(681.38 KB) Publisher Site Additional Information: full citation, abstract, references, citings

We present some variations affecting the association measure and thresholding on a technique for learning Selectional Restrictions from on-line corpora. It uses a wide-coverage noun taxonomy and a statistical measure to generalize the appropriate semantic classes. Evaluation measures for the Selectional Restrictions learning task are discussed. Finally, an experimental evaluation of these variations is reported.

Keywords: computational lexicography, corpus-based language modeling

14 The SMART lab report

Mike Lesk, Donna Harman, Edward A. Fox, Harry Wu, Chris Buckley

April 1997 **ACM SIGIR Forum**, Volume 31 Issue 1

Full text available: pdf(1.65 MB) Additional Information: full citation, index terms

15 Session: Complementing WordNet with Roget's and corpus-based thesauri for information retrieval

Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka

June 1999 **Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics**

Full text available: pdf(647.88 KB) Publisher Site Additional Information: full citation, abstract, references

This paper proposes a method to overcome the drawbacks of WordNet when applied to information retrieval by complementing it with Roget's thesaurus and corpus-derived thesauri. Words and relations which are not included in WordNet can be found in the corpus-

derived thesauri. Effects of polysemy can be minimized with weighting method considering all query terms and all of the thesauri. Experimental results show that our method enhances information retrieval performance significantly.

16 Special issue on word sense disambiguation: Using corpus statistics and WordNet relations for sense identification

Claudia Leacock, George A. Miller, Martin Chodorow

March 1998 **Computational Linguistics**, Volume 24 Issue 1

Full text available: pdf(1.35 MB) Additional Information: full citation, abstract, references, citings

Publisher Site

Corpus-based approaches to word sense identification have flexibility and generality but suffer from a knowledge acquisition bottleneck. We show how knowledge-based techniques can be used to open the bottleneck by automatically locating training corpora. We describe a statistical classifier that combines topical context with local cues to identify a word sense. The classifier is used to disambiguate a noun, a verb, and an adjective. A knowledge base in the form of WordNet's lexical relations is ...

17 Qusetion answering: Structured use of external knowledge for event-based open domain question answering

Hui Yang, Tat-Seng Chua, Shuguang Wang, Chun-Keat Koh

July 2003 **Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval**

Full text available: pdf(287.93 KB) Additional Information: full citation, abstract, references, citings, index terms

One of the major problems in question answering (QA) is that the queries are either too brief or often do not contain most relevant terms in the target corpus. In order to overcome this problem, our earlier work integrates external knowledge extracted from the Web and WordNet to perform Event-based QA on the TREC-11 task. This paper extends our approach to perform event-based QA by uncovering the structure within the external knowledge. The knowledge structure loosely models different facets of ...

Keywords: event-based QA, query formulation, question answering

18 The use of phrases and structured queries in information retrieval

W. Bruce Croft, Howard R. Turtle, David D. Lewis

September 1991 **Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(1.35 MB) Additional Information: full citation, references, citings, index terms

19 Efficiency and scaling: Hourly analysis of a very large topically categorized web query log

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder

July 2004 **Proceedings of the 27th annual international conference on Research and development in information retrieval**

Full text available: pdf(285.93 KB) Additional Information: full citation, abstract, references, index terms

We review a query log of hundreds of millions of queries that constitute the total query traffic for an entire week of a general-purpose commercial web search service. Previously, query logs have been studied from a single, cumulative view. In contrast, our analysis shows changes in popularity and uniqueness of topically categorized queries across the hours of the day. We examine query traffic on an hourly basis by matching it against lists of

queries that have been topically pre-categorized by ...

Keywords: query log analysis, web search

20 Special issue on word sense disambiguation: Disambiguating highly ambiguous words 
Geoffrey Towell, Ellen M. Voorhees
March 1998 **Computational Linguistics**, Volume 24 Issue 1

Full text available: pdf(1.48 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

A word sense disambiguator that is able to distinguish among the many senses of common words that are found in general-purpose, broad-coverage lexicons would be useful. For example, experiments have shown that, given accurate sense disambiguation, the lexical relations encoded in lexicons such as WordNet can be exploited to improve the effectiveness of information retrieval systems. This paper describes a classifier whose accuracy may be sufficient for such a purpose. The classifier combines the ...

Results 1 - 20 of 200 Result page: **1** 2 3 4 5 6 7 8 9 10 next

Terms used **phrase frquency** and **phrase co occurrence** and **taxonomy** and **normalization**

Found **7,683** of **147,060**

Sort results by: relevance

Display results: expanded form

Save results to a Binder

Search Tips

☐ Open results in a new window

Try an Advanced Search

Try this search in The ACM Guide

Results 21 - 40 of 200 Result page: previous 1 **2** 3 4 5 6 7 8 9 10 next

Best 200 shown

Relevance scale

21 Hypertext data mining (tutorial AM-1)

Soumen Chakrabarti

August 2000 **Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining**

Full text available: pdf(1.08 MB) Additional Information: full citation, index terms

22 Student session paper: Corpus-based identification of non-anaphoric noun phrases

David L. Bean, Ellen Riloff

June 1999 **Proceedings of the 37th conference on Association for Computational Linguistics**

Full text available: pdf(673.88 KB) Publisher Site Additional Information: full citation, abstract, references, citings

Coreference resolution involves finding antecedents for anaphoric discourse entities, such as definite noun phrases. But many definite noun phrases are not anaphoric because their meaning can be understood from general world knowledge (e.g., "the White House" or "the news media"). We have developed a corpus-based algorithm for automatically identifying definite noun phrases that are non-anaphoric, which has the potential to improve the efficiency and accuracy of coreference resolution systems. O ...

23 Empirical studies on the disambiguation of cue phrases

Julia Hirschberg, Diane Litman

September 1993 **Computational Linguistics**, Volume 19 Issue 3

Full text available: pdf(2.05 MB) Publisher Site Additional Information: full citation, abstract, references, citings

Cue phrases are linguistic expressions such as now and well that function as explicit indicators of the structure of a discourse. For example, now may signal the beginning of a subtopic or a return to a previous topic, while well may mark subsequent material as a response to prior material, or as an explanatory comment. However, while cue phrases may convey discourse structure, each also has one or more alternate uses. While incidentally may be used **sententially** as an adverbial, for examp ...

24 An evaluation of phrasal and clustered representations on a text categorization task

David D. Lewis
June 1992 **Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(1.22 MB) Additional Information: full citation, abstract, references, citings, index terms

Syntactic phrase indexing and term clustering have been widely explored as text representation techniques for text retrieval. In this paper we study the properties of phrasal and clustered indexing languages on a text categorization task, enabling us to study their properties in isolation from query interpretation issues. We show that optimal effectiveness occurs when using only a small proportion of the indexing terms available, and that effectiveness peaks at a higher feature set size and ...

25 Interacting through different modalities: Visual display, pointing, and natural language: the power of multimodal interaction
Antonella De Angeli, Walter Gerbino, Giulia Cassano, Daniela Petrelli
May 1998 **Proceedings of the working conference on Advanced visual interfaces**

Full text available: pdf(1.56 MB) Additional Information: full citation, abstract, references

This paper examines user behavior during multimodal human-computer interaction (HCI). It discusses how pointing, natural language, and graphical layout should be integrated to enhance the usability of multimodal systems. Two experiments were run to study simulated systems capable of understanding written natural language and mouse-supported pointing gestures. Results allowed to: (a) develop a taxonomy of communication acts aimed at identifying targets; (b) determine the conditions under which sp ...

Keywords: cross-modal integration, referent identification strategies

26 Ada and the evolution of software engineering
Neal Coulter, Ira Monarch, Suresh Konda, Marvin Carr
November 1995 **Proceedings of the conference on TRI-Ada '95: Ada's role in global markets: solutions for a changing complex world**

Full text available: pdf(1.24 MB) Additional Information: full citation, references

27 Theory of keyblock-based image retrieval
April 2002 **ACM Transactions on Information Systems (TOIS)**, Volume 20 Issue 2

Full text available: pdf(2.14 MB) Additional Information: full citation, abstract, references, index terms, review

The success of text-based retrieval motivates us to investigate analogous techniques which can support the querying and browsing of image data. However, images differ significantly from text both syntactically and semantically in their mode of representing and expressing information. Thus, the generalization of information retrieval from the text domain to the image domain is non-trivial. This paper presents a framework for information retrieval in the image domain which supports content-based q ...

Keywords: clustering, codebook, content-based image retrieval, keyblock

28 A phrase-structured grammatical framework for transportable natural language processing
Bruce W. Ballard, Nancy L. Tinkham
April 1984 **Computational Linguistics**, Volume 10 Issue 2

Full text available: pdf(1.51 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

We present methods of dealing with the syntactic problems that arise in the construction of natural language processors that seek to allow users, as opposed to computational linguists, to customize an interface to operate with a new domain of data. In particular, we describe a *grammatical formalism*, based on augmented phrase-structure rules, which allows a parser to perform many important domain-specific disambiguations by reference to a pre-defined grammar and a collection of auxiliary f ...

29 Browsing in digital libraries: a phrase-based approach
Craig G. Nevill-Manning, Ian H. Witten, Gordon W. Paynter
July 1997 **Proceedings of the second ACM international conference on Digital libraries**

Full text available: pdf(1.22 MB) Additional Information: full citation, references, citings, index terms

30 Computer Evaluation of Indexing and Text Processing
G. Salton, M. E. Lesk
January 1968 **Journal of the ACM (JACM)**, Volume 15 Issue 1

Full text available: pdf(2.19 MB) Additional Information: full citation, abstract, references, citings, index terms

Automatic indexing methods are evaluated and design criteria for modern information systems are derived.

31 Student session paper: Measures of distributional similarity
Lillian Lee
June 1999 **Proceedings of the 37th conference on Association for Computational Linguistics**

Full text available: pdf(705.14 KB) Additional Information: full citation, abstract, references, citings
Publisher Site

We study distributional similarity measures for the purpose of improving probability estimation for unseen cooccurrences. Our contributions are three-fold: an empirical comparison of a broad range of measures; a classification of similarity functions based on the information that they incorporate; and the introduction of a novel function that is superior at evaluating potential proxy distributions.

32 SCAN: designing and evaluating user interfaces to support retrieval from speech archives
Steve Whittaker, Julia Hirschberg, John Choi, Don Hindle, Fernando Pereira, Amit Singhal
August 1999 **Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(161.15 KB) Additional Information: full citation, references, citings, index terms

Keywords: comparing interfaces for information access, field/empirical studies of the information seeking process, speech indexing and retrieval, user studies

33 Measures of distributional similarity
Lillian Lee
June 1999 **Proceedings of the 37th conference on Association for Computational Linguistics**

Full text available: pdf(705.14 KB) Additional Information: full citation, abstract, references

We study distributional similarity measures for the purpose of improving probability estimation for unseen cooccurrences. Our contributions are three-fold: an empirical comparison of a broad range of measures; a classification of similarity functions based on the information that they incorporate; and the introduction of a novel function that is superior at evaluating potential proxy distributions.

34 Exploring the similarity space

Justin Zobel, Alistair Moffat
April 1998 **ACM SIGIR Forum**, Volume 32 Issue 1

Full text available: pdf(1.23 MB) Additional Information: full citation, abstract, citings, index terms

Ranked queries are used to locate relevant documents in text databases. In a ranked query a list of terms is specified, then the documents that most closely match the query are returned---in decreasing order of similarity---as answers. Crucial to the efficacy of ranked querying is the use of a similarity heuristic, a mechanism that assigns a numeric score indicating how closely a document and the query match. In this note we explore and categorise a range of similarity heuristics described in th ...

35 Information extraction: Web-scale information extraction in knowitall: (preliminary results)

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates
May 2004 **Proceedings of the 13th international conference on World Wide Web**

Full text available: pdf(171.42 KB) Additional Information: full citation, abstract, references, citings, index terms

Manually querying search engines in order to accumulate a large bodyof factual information is a tedious, error-prone process of piecemealsearch. Search engines retrieve and rank potentially relevantdocuments for human perusal, but do not extract facts, assessconfidence, or fuse information from multiple documents. This paperintroduces KnowItAll, a system that aims to automate the tedious process ofextracting large collections of facts from the web in an autonomous,domain-independent, and scalabl ...

Keywords: information extraction, mutual information, pmi, search

36 Query result processing: A hierarchical monothetic document clustering algorithm for summarization and browsing search results

Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, Raghu Krishnapuram
May 2004 **Proceedings of the 13th international conference on World Wide Web**

Full text available: pdf(446.94 KB) Additional Information: full citation, abstract, references, index terms

Organizing Web search results into a hierarchy of topics and sub-topics facilitates browsing the collection and locating results of interest. In this paper, we propose a new hierarchical monothetic clustering algorithm to build a topic hierarchy for a collection of search results retrieved in response to a query. At every level of the hierarchy, the new algorithm progressively identifies topics in a way that maximizes the coverage while maintaining distinctiveness of the topics. We refer the pro ...

Keywords: automatic taxonomy generation, clustering, data mining, search, summarization

37 Special issue on word sense disambiguation: Introduction to the special issue on word sense disambiguation: the state of the art
Nancy Ide, Jean Véronis
March 1998 **Computational Linguistics**, Volume 24 Issue 1

Full text available: pdf(3.44 MB) Additional Information: full citation, references, citings
Publisher Site

38 Building effective queries in natural language information retrieval
Tomek Strzalkowski, Fang Lin, Jose Perez-Carballo, Jin Wang
March 1997 **Proceedings of the fifth conference on Applied natural language processing**

Full text available: pdf(771.03 KB) Additional Information: full citation, abstract, references, citings
Publisher Site

In this paper we report on our natural language information retrieval (NLIR) project as related to the recently concluded 5th Text Retrieval Conference (TREC-5). The main thrust of this project is to use natural language processing techniques to enhance the effectiveness of full-text document retrieval. One of our goals was to demonstrate that robust if relatively shallow NLP can help to derive a better representation of text documents for statistical search. Recently, we have turned our attenti ...

39 Special issue on using large corpora: II: Lexical semantic techniques for corpus analysis
James Pustejovsky, Peter Anick, Sabine Bergler
June 1993 **Computational Linguistics**, Volume 19 Issue 2

Full text available: pdf(1.90 MB) Additional Information: full citation, abstract, references, citings
Publisher Site

In this paper we outline a research program for computational linguistics, making extensive use of text corpora. We demonstrate how a semantic framework for lexical knowledge can suggest richer relationships among words in text beyond that of simple co-occurrence. The work suggests how linguistic phenomena such as metonymy and polysemy might be exploitable for semantic tagging of lexical items. Unlike with purely statistical collocational analyses, the framework of a semantic theory allows the a ...

40 The rhetorical parsing of unrestricted texts: a surface-based approach
Daniel Marcu
September 2000 **Computational Linguistics**, Volume 26 Issue 3

Full text available: pdf(3.87 MB) Additional Information: full citation, abstract, references
Publisher Site

Coherent texts are not just simple sequences of clauses and sentences, but rather complex artifacts that have highly elaborate rhetorical structure. This paper explores the extent to which well-formed rhetorical structures can be automatically derived by means of surface-form-based algorithms. These algorithms identify discourse usages of cue phrases and break sentences into clauses, hypothesize rhetorical relations that hold among textual units, and produce valid rhetorical structure trees for ...

Terms used **phrase frquency** and **phrase co occurrence** and **taxonomy** and **normalization**

Found **7,683** of **147,060**

Sort results by relevance

Display results expanded form

Save results to a Binder

Search Tips

Open results in a new window

Try an Advanced Search

Try this search in The ACM Guide

Results 41 - 60 of 200 Result page: previous 1 2 **3** 4 5 6 7 8 9 10 next

Best 200 shown

Relevance scale

41 High performance question/answering

Marius A. Pasca, Sandra M. Harabagiu

September 2001 **Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(256.26 KB) Additional Information: full citation, abstract, references, citings, index terms

In this paper we present the features of a Question/Answering (Q/A) system that had unparalleled performance in the TREC-9 evaluations. We explain the accuracy of our system through the unique characteristics of its architecture: (1) usage of a wide-coverage answer type taxonomy; (2) repeated passage retrieval; (3) lexico-semantic feedback loops; (4) extraction of the answers based on machine learning techniques; and (5) answer caching. Experimental results show the effects of each feature ...

42 Discovery procedures for sublanguage selectional patterns: initial experiments

Ralph Grishman, Lynette Hirschman, Ngo Thanh Nhan

July 1986 **Computational Linguistics**, Volume 12 Issue 3

Full text available: pdf(1.09 MB) Publisher Site Additional Information: full citation, abstract, references, citings

Selectional constraints specify, for a particular domain, the combinations of semantic classes acceptable in subject-verb-object relationships and other syntactic structures. These constraints are important in blocking incorrect analyses in natural language processing systems. However, these constraints are domain-specific and hence must be developed anew when a system is ported to a new domain. A discovery procedure for selectional constraints is therefore essential in enhancing the portability ...

43 Cross-language multimedia information retrieval

Sharon Flank

April 2000 **Proceedings of the sixth conference on Applied natural language processing**

Full text available: pdf(1.01 MB) Publisher Site Additional Information: full citation, abstract, references

Simple measures can achieve high-accuracy cross-language retrieval in carefully chosen applications. Image retrieval is one of those applications, with results ranging from 68% of human translator performance for German, to 100% for French.

44 Structural disambiguation based on reliable estimation of strength of association
Haodong Wu, Eduardo de Paiva Alves, Teiji Furugori
August 1998

Full text available: pdf(574.67 KB) Publisher Site Additional Information: full citation, abstract, references

This paper proposes a new class-based method to estimate the strength of association in word co-occurrence for the purpose of structural disambiguation. To deal with sparseness of data, we use a conceptual dictionary as the source for acquiring upper classes of the words related in the co-occurrence, and then use t-scores to determine a pair of classes to be employed for calculating the strength of association. We have applied our method to determining dependency relations in Japanese and prepos ...

45 Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology
Yu-Sheng Lai, Chung-Hsien Wu
March 2002 **ACM Transactions on Asian Language Information Processing (TALIP)**, Volume 1 Issue 1

Full text available: pdf(920.43 KB) Additional Information: full citation, abstract, references, index terms

In this article, an approach based on unknown words is proposed for meaningful term extraction and discriminative term selection in text categorization. For meaningful term extraction, a phrase-like unit (PLU)-based likelihood ratio is proposed to estimate the likelihood that a word sequence is an unknown word. On the other hand, a discriminative measure is proposed for term selection and is combined with the PLU-based likelihood ratio to determine the text category. We conducted several experim ...

Keywords: AC-machine, dimensionality reduction, discriminability, discriminative term selection, inconsistency problem, meaningful term extraction, n-gram, phrase-like unit, sparse data problem, term adaptation, term purification, text categorization, text indexing, unknown word detection, vector space modeling

46 Natural language information retrieval in digital libraries
Tomek Strzalkowski, Jose Perez-Carballo, Mihnea Marinescu
April 1996 **Proceedings of the first ACM international conference on Digital libraries**

Full text available: pdf(1.03 MB) Additional Information: full citation, references, index terms

47 Detecting topical events in digital video
Tanveer Syeda-Mahmood, S. Srinivasan
October 2000 **Proceedings of the eighth ACM international conference on Multimedia**

Full text available: pdf(1.04 MB) Additional Information: full citation, abstract, references, citings, index terms

The detection of events is essential to high-level semantic querying of video databases. It is also a very challenging problem requiring the detection and integration of evidence for an event available in multiple information modalities, such as audio, video and language. This paper focuses on the detection of specific types of events, namely, topic of discussion events that occur in classroom/lecture environments. Specifically, we present a query-driven approach to the detection of topic of ...

Keywords: multi-modal fusion, query-driven topic detection, slide detection, topic of discussion events, topical audio events

48 Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies
Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, Prabhakar Raghavan
August 1998 **The VLDB Journal — The International Journal on Very Large Data Bases**, Volume 7 Issue 3
Full text available: pdf(281.37 KB) Additional Information: full citation, abstract, citings, index terms

We explore how to organize large text databases hierarchically by topic to aid better searching, browsing and filtering. Many corpora, such as internet directories, digital libraries, and patent databases are manually organized into topic hierarchies, also called *taxonomies*. Similar to indices for relational data, taxonomies make search and access more efficient. However, the exponential growth in the volume of on-line textual information makes it nearly impossible to maintain such taxono ...

49 Use of syntactic context to produce term association lists for text retrieval
Gregory Grefenstette
June 1992 **Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval**
Full text available: pdf(714.72 KB) Additional Information: full citation, abstract, references, citings, index terms

One aspect of world knowledge essential to information retrieval is knowing when two words are related. Knowing word relatedness allows a system given a user's query terms to retrieve relevant documents not containing those exact terms. Two words can be said to be related if they appear in the same contexts Document co-occurrence gives a measure of word relatedness that has proved to be too rough to be useful. The relatively recent apparition of on-line dictionaries and robust and rapid par ...

50 An application of plausible reasoning to information retrieval
Farhad Oroumchian, Robert N. Oddy
August 1996 **Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval**
Full text available: pdf(808.70 KB) Additional Information: full citation, references, index terms

51 On the application of syntactic methodologies in automatic text analysis
G. Salton, M. Smith
May 1989 **ACM SIGIR Forum , Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval**, Volume 23 Issue 1-2
Full text available: pdf(1.09 MB) Additional Information: full citation, abstract, references, citings, index terms

This study summarizes various linguistic approaches proposed for document analysis in information retrieval environments. Included are standard syntactic methods to generate complex content identifiers, and the use of semantic know-how obtained from machine-readable dictionaries and from specially constructed knowledge bases. A particular syntactic analysis methodology is also outlined and its usefulness for the automatic construction of book indexes is examined.

52 Query clustering using user logs
January 2002 **ACM Transactions on Information Systems (TOIS)**, Volume 20 Issue 1
Full text available: pdf(1.31 MB) Additional Information: full citation, abstract, references, citings, index terms, review

Query clustering is a process used to discover frequently asked questions or most popular topics on a search engine. This process is crucial for search engines based on question-answering. Because of the short lengths of queries, approaches based on keywords are not suitable for query clustering. This paper describes a new query clustering method that makes use of user logs which allow us to identify the documents the users have selected for a query. The similarity between two queries may be ded ...

Keywords: Query clustering, search engine, user log, web data mining

53 Text categorization and retrieval: Robust text processing in automated information retrieval

Tomek Strzalkowski

October 1994 **Proceedings of the fourth conference on Applied natural language processing**

Full text available: pdf(593.70 KB) Publisher Site Additional Information: full citation, abstract, references, citings

We report on the results of a series of experiments with a prototype text retrieval system which uses relatively advanced natural language processing techniques in order to enhance the effectiveness of statistical document retrieval. In this paper we show that large-scale natural language processing (hundreds of millions of words and more) is not only required for a better retrieval, but it is also doable, given appropriate resources. In particular, we demonstrate that the use of syntactic compo ...

54 Special issue on using large corpora: I: Generalized probabilistic LR parsing of natural language (Corpora) with unification-based grammars

Ted Briscoe, John Carroll

March 1993 **Computational Linguistics**, Volume 19 Issue 1

Full text available: pdf(2.62 MB) Publisher Site Additional Information: full citation, abstract, references, citings

We describe work toward the construction of a very wide-coverage probabilistic parsing system for natural language (NL), based on LR parsing techniques. The system is intended to rank the large number of syntactic analyses produced by NL grammars according to the frequency of occurrence of the individual rules deployed in each analysis. We discuss a fully automatic procedure for constructing an LR parse table from a unification-based grammar formalism, and consider the suitability of alternative ...

55 Learning search engine specific query transformations for question answering

Eugene Agichtein, Steve Lawrence, Luis Gravano

April 2001 **Proceedings of the tenth international conference on World Wide Web**

Full text available: pdf(205.68 KB) Additional Information: full citation, references, citings, index terms

Keywords: information retrieval, query expansion, question answering, web search

56 Semantics III: On the use of term associations in automatic information retrieval

Gerard Salton

August 1986 **Proceedings of the 11th coference on Computational linguistics**

Full text available: pdf(622.43 KB) Additional Information: full citation, abstract, references

It has been recognized that single words extracted from natural language texts are not

always useful for the representation of information content. Associated or related terms. and comples content identifiers derived from thesauruses and knowledge bases, or constructed by automatic word grouping techniques, have therefore been proposed for text identification purposes.The area of associative content analysis and information retrievl is reviewed in this study. The available experimental evidence ...

57 Information retrieval using robust natural language processing

Tomek Strzalkowski, Barbara Vauthey

June 1992 **Proceedings of the 30th conference on Association for Computational Linguistics**

Full text available: pdf(772.67 KB) Publisher Site Additional Information: full citation, abstract, references, citings

We developed a prototype information retrieval system which uses advanced natural language processing techniques to enhance the effectiveness of traditional key-word based document retrieval. The backbone of our system is a statistical retrieval engine which performs automated indexing of documents, then search and ranking in response to user queries. This core architecture is augmented with advanced natural language processing tools which are both robust and efficient. In early experiments, the ...

58 Special issue on using large corpora: I: Introduction to the special issue on computational linguistics using large corpora

Kenneth W. Church, Robert L. Mercer

March 1993 **Computational Linguistics**, Volume 19 Issue 1

Full text available: pdf(1.80 MB) Publisher Site Additional Information: full citation, references, citings

59 Measuring praise and criticism: Inference of semantic orientation from association

Peter D. Turney, Michael L. Littman

October 2003 **ACM Transactions on Information Systems (TOIS)**, Volume 21 Issue 4

Full text available: pdf(640.81 KB) Additional Information: full citation, abstract, references, index terms

The evaluative character of a word is called its *semantic orientation*. Positive semantic orientation indicates praise (e.g., "honest", "intrepid") and negative semantic orientation indicates criticism (e.g., "disturbing", "superfluous"). Semantic orientation varies in both direction (positive or negative) and degree (mild to strong). An automated system for measuring semantic orientation would have application in text classification, text filtering, tracking opinions in online discussions ...

Keywords: latent semantic analysis, mutual information, semantic association, semantic orientation, text classification, text mining, unsupervised learning, web mining

60 Recent trends in automatic information retrieval

Gerard Salton

September 1986 **Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval**

Full text available: pdf(1.05 MB) Additional Information: full citation, abstract, references, citings

Substantial successes were achieved in the early years in automatic indexing and retrieval using single term indexing theories with term weight assignments based on frequency considerations. The development of more refined indexing systems using thesaurus aids and automatically constructed term association maps changed the retrieval effectiveness only slightly. The recent introduction of the relevance concept in the form of probabilistic retrieval

models provided a firm basis for term weigh ...

Results 41 - 60 of 200 Result page: previous 1 2 **3** 4 5 6 7 8 9 10 next

09360,154

L21: Entry 2 of 4 File: USPT Nov 18, 2003

DOCUMENT-IDENTIFIER: US 6651058 B1
TITLE: System and method of automatic discovery of terms in a document that are relevant to a given target topic

Abstract Text (1):
A computer program product is provided as an automatic mining system to discover terms that are relevant to a given target topic from a large databases of unstructured information such as the World Wide Web. The operation of the automatic mining system is performed in three stages: The first stage is carried out by a new terms discoverer for discovering the terms in a document, the second stage is carried out by a candidate terms discoverer for discovering potentially relevant terms, and the third stage is carried out by a relevant terms discoverer for refining or testing the discovered relevance to filter false relevance. The new terms discoverer includes a system for the automatic mining of patterns and relations, a system for the automatic mining of new relationships, and a system for selecting new terms from relations. In one embodiment, the system for the automatic mining of patterns and relations identifies a set of related terms on the WWW with a high degree of confidence, using a duality concept, and includes a terms database and two identifiers: a relation identifier and a pattern identifier. The system for the automatic mining of new relationships includes a database a knowledge module and a statistics module. The knowledge module includes a stemming unit, a synonym check unit, and a domain knowledge check unit. The candidate terms discoverer includes a metadata extractor, a document vector module, an association module, a filtering module, and a database. The relevant terms discoverer includes a stop word filter and a system for the automatic construction of generalization--specialization hierarchy of terms comprised of a terms database, an augmentation module, a generalization detection module, and a hierarchy database.

Brief Summary Text (2):
The present invention relates to the field of data mining, and particularly to a software system and associated methods for automatically discovering terms that are relevant to a given target topic from a large databases of unstructured information such as the World Wide Web (WWW). More specifically, the present invention relates to the automatic and iterative recognition of relevant terms by association mining and refinement of co-occurrences.

Brief Summary Text (4):
The World Wide Web (WWW) is a vast and open communications network where computer users can access available data, digitally encoded documents, books, pictures, and sounds. With the explosive growth and diversity of WWW authors, published information is oftentimes unstructured and widely scattered. Although search engines play an important role in furnishing desired information to the end users, the organization of the information lacks structure and consistency. Web spiders crawl web pages and index them to serve the search engines. As the web spiders visit web pages, they could look for, and learn pieces of information that would otherwise remain undetected.

Brief Summary Text (5):
Current search engines are designed to identify pages with specific phrases and offer limited search capabilities. For example, search engines cannot search for phrases that relate in a particular way, such as books and authors. Bibliometrics involves the study of the world of authorship and citations. It measures the co-citation strength, which is a measure of the similarity between two technical papers on the basis of their common citations. Statistical techniques are used to compute this measures. In typical bibliometric situations the citations and authorship are explicit and do not need to be mined. One of the limitations of the

bibliometrics is that it cannot be used to extract buried information in the text.

Brief Summary Text (6):
Exemplary bibliometric studies are reported in: R. Larson, "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace," Technical report, School of Information Management and Systems, University of California, Berkeley, 1996. hftp://sherlock.sims.berkeley.edu/docs/asis96/asis96.html; K. McCain, "Mapping Authors in Intellectual Space: A technical Overview," Journal of the American Society for Information Science, 41(6):433-443, 1990. A Dual Iterative Pattern Relation Expansion (DIPRE) method that addresses the problem of extracting (author, book) relationships from the web is described in S. Brin, "Extracting Patterns and Relations from the World Wide Web," WebDB, Valencia, Spain, 1998.

Brief Summary Text (11):
Exemplary work in scalable data mining technology, is described in the following references: R. Agrawal et al., "Mining Association Rules Between Sets of Items in Large Databases, Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993; R. Agrawal et al., "Fast Algorithms for Mining Association Rules," Proc. of the 20th Int'l Conference on VLDB, Santiago, Chile, September 1994; and S. Brin, "Extracting Patterns and Relations from the World Wide Web," WebDB, Valencia, Spain, 1998, supra. Such work has been successfully applied to identify co-occurring patterns in many real world problems including market basket analysis, cross-marketing, store layout, and customer segmentation based on buying patterns.

Brief Summary Text (15):
In accordance with the present invention, a computer program product is provided as an automatic mining system to discover terms that are relevant to a given target topic from a large databases of unstructured information such as the World Wide Web (WWW). The system and methods enable the automatic and iterative recognition of relevant terms by association mining and refinement of co-occurrences.

Brief Summary Text (18):
The system for the automatic mining of new relationships enables the discovery of new relationships by association mining and refinement of co-occurrences, using automatic and iterative recognition of new binary relations through phrases that embody related pairs. The system for the automatic mining of new relationships is comprised of a database a knowledge module and a statistics module. In one embodiment, the knowledge module includes one or more of the following units: a stemming unit, a synonym check unit, and a domain knowledge check unit. New terms are obtained from relations discovered by the system for automatic mining of patterns and relations of the same kind by selecting an item (or a column) of a pair.

Brief Summary Text (19):
The candidate terms discoverer is comprised of a metadata extractor, a document vector module, an association module, a filtering module, and a database for storing the mined sets of relevant terms. The relevant terms discoverer includes a stop word filter and a system for the automatic construction of generalization-specialization hierarchy of terms. The system for the automatic construction of generalization-specialization hierarchy of terms includes a terms database, an augmentation module, a generalization detection module, and a hierarchy database.

Detailed Description Text (24):
Using the terms mined by the new terms discoverer 90, the candidate terms discoverer 100 identifies potentially relevant terms by the frequency of their co-occurrence. The relevant terms discoverer 110 filters the terms mined by the candidate terms discoverer 100 to determine the accuracy of these terms and their meaningful (or close) relevance to the target concept. This is accomplished by eliminating non-meaningful co-occurring terms. For example, banana is a "major" or frequent term that occurs in a significant number of grocery shopping carts (i.e., banana is frequently purchased), and has no particular relevance to other co-occurring terms (i.e., to other purchased grocery items). As a result, the relevant terms discoverer 110 can eliminate banana from the list of closely relevant terms. The terms resulting from the new terms discoverer 90 are considered to be relevant and are stored in the relevant terms database

130.

Detailed Description Text (25):
Having described the main components of the automatic mining system 10, its operation will now be further explained in connection with FIGS. 3 through 8. The first stage of the operation is carried out by the new terms discoverer 90 which is illustrated in FIGS. 3, 4 and 5. The new terms discoverer 90 defines the terms that need to be examined for their relevance to a target topic (or concept). As used herein a target topic can be defined as a description of the cluster of the topic's instances in a database. For a given target topic and a database of HTML documents discovering a relevant topic is to discover a topic whose cluster significantly overlaps with the target topic's cluster. In other words, a significant number of the relevant topic's instances belong to the target topic's cluster. The significance is determined by a user-defined threshold. In the illustration described herein, relevance is defined in terms of co-occurrence of terms.

Detailed Description Text (34):
FIG. 6 illustrates an exemplary candidate terms discoverer 100, which is comprised of a metadata extractor 220, a document vector module 230, an association module 240, and a filtering module 250. The candidate terms discoverer 100 further includes a database 260 for storing the mined sets of relevant terms. The set of relevant terms is continuously and iteratively broadened by the candidate terms discoverer 100.

Detailed Description Text (35):
The metadata extractor 220 identifies all the hypertext link metadata in the document di. Whereupon, the document vector module 230 creates a document vector for each document d.sub.i. In a preferred embodiment, the document vector module 230 does not list duplicate terms or the frequency of occurrence of all the terms. Rather, the association module 240 measures the number of documents that contain the terms, regardless of the frequency of occurrence of the terms within a single document d.sub.i. Such measurement enables the association module 240 to perform the necessary statistical analyses.

Detailed Description Text (37):
where t is a term and T is a topic term, rel (t{character pullout}T) is a relevance metric of the association rule such as support of confidence, and c is a user-specified threshold. From R, a set of candidate terms CT is extracted as follows:

Detailed Description Text (41):
The system 275 for the automatic construction of generalization-specialization hierarchy of terms builds a relevance model based upon a generalization-specialization hierarchy. Initially, starting with only the target topic, the relevance model is progressively constructed by new terms and the generalization relationship between the new terms and the existing terms in the relevance model. A generalization relationship between a new candidate term (Ct) and a term in the relevance model (rt), is added to the relevance model if the taxonomy declares the candidate term (ct) to be a specialization of the term in the relevance model (rt).

Detailed Description Text (42):
In addition to the specialization of relevant terms, the Least General Generalization (LGG) terms in the relevance model are added to the relevance model, if the LGG is not overgeneralized. A generalized term by LGG is said to be overgeneralized, if the co-occurrence between the generalized term and its specialization in the relevance model is below a predetermined threshold. As an example, in a random sample set (S) of documents the co-occurrence of the LGG and a term (rt) in the relevance model that is a specialization of the LGG is measured by the joint probability of the LGG and the term (rt). If the joint probability of the LGG and the relevance model in the set (S) is less than a user provided threshold (t), i.e. pS(LGG, rt)<t, the LGG is overgeneralized, and the LGG is not added to the relevance model. Based on this relevance model, a candidate term is determined to be relevant to the target topic if the candidate term is a specialization of a term in the relevance model.

Detailed Description Text (45):
One function of the augmentation module 282 is to update the set of terms knowing the terms

stored in the terms database 280. This feature is implemented by a generalization technique such as the "Least General Generalization" or LGG model. The generalization detection module maps the LGG sets that are stored in the terms database 280 and the LGG terms that are derived by the augmentation module 282, updates the set of edges (or hierarchical or generalization relationships), and derives a generalization hierarchy. In operation, the system 275 begins with no predefined taxonomy of the terms, and the LGG model derives a generalization hierarchy, modeled as a Directed Acyclic Graph (DAG), from the set of terms. The generalization hierarchy maps the generalization and specialization relationships between the terms. A more complete description of the system 140 can be found in U.S. patent application Ser. No. 09/440,203, is now pending titled "System and Method for the Automatic Construction of Generalization-Specialization Hierarchy of Terms", which is incorporated herein by reference.

Current US Original Classification (1):
707/6

Current US Cross Reference Classification (3):
707/100

Current US Cross Reference Classification (4):
707/102

Current US Cross Reference Classification (5):
707/3

Current US Cross Reference Classification (6):
707/5

Other Reference Publication (5):
S. Brin, "Extracting Patterns and Relations from the World Wide Web," WebDB, Valencia, Spain, 1998.

Other Reference Publication (16):
S. Soderland. Learning to Extract Text-based Information from the World Wide Web, American Association for Artificial Intelligence (www.aaai.org), pp. 251-254.

CLAIMS:

8. The system according to claim 1, wherein the candidate terms discoverer includes a metadata extractor.

11. The system according to claim 9, wherein the metadata extractor identifies hypertext link metadata in the document, and the document vector module creates a document vector for each document.

17. The system according to claim 15, wherein the candidate terms discoverer includes a metadata extractor, a document vector module, an association module, a filtering module, and a database for storing relevant terms; and wherein the system for the automatic construction of the generalization hierarchy of terms includes an augmentation module, a generalization detection module, and a hierarchy database.

Previous Doc Next Doc Go to Doc#

L19: Entry 5 of 9 File: USPT Aug 27, 2002

DOCUMENT-IDENTIFIER: US 6442545 B1
TITLE: Term-level text with mining with taxonomies

Abstract Text (1):
A method for mining in a database including documents, the documents including text. The method includes providing a taxonomy of taxonomy terms, and mining the documents responsive to the taxonomy to discover a relationship between a set of one or more selected words and at least one of the taxonomy terms. The method also includes analyzing occurrences of the relationship over a plurality of the documents to extract information relating to the at least one taxonomy term.

Brief Summary Text (2):
The present invention relates generally to extraction of information from databases, and specifically to text mining in unstructured databases.

Brief Summary Text (5):
A paper entitled "Technology Text Mining, Turning Information Into Knowledge: A White Paper from IBM," edited by Daniel Tkach, Feb. 17, 1998, which is incorporated herein by reference, describes a program called IBM Intelligent Miner for Text, which extracts terms from unstructured text. "Terms," in the context of the present application, are single words, or short strings of highly-related, linked words, such as "Biotechnology," "New York Stock Exchange," "Free market," or "Health programs." "Term extraction," in the context of the present application, refers to the process of finding terms in a document that have relevance to the content of the document.

Brief Summary Text (6):
InQuery 5.0, produced by Sovereign Hill Software, uses term extraction to identify names of companies and people in one or more documents. The extracted terms are used to enable a search engine to find desired documents responsive to a user's query.

Brief Summary Text (7):
A paper entitled "Text Mining at the Term Level," by Feldman et al., Proceedings of the 1998 Workshop on Knowledge Discovery in Databases, August, 1998, which is incorporated herein by reference, the authors of which are the inventors of the present invention, describes a method for extracting terms from a document in a database, filtering out unimportant terms, and subsequently performing text mining in the database. "Text mining," in the context of the present application, refers to a substantially automated process of extracting useful information from a collection of textual data.

Brief Summary Text (12):
It is yet a further object of some aspects of the present invention to provide improved methods for extracting information from multiple documents in a database.

Brief Summary Text (13):
In preferred embodiments of the present invention, a system for mining text in a database comprises a memory, which stores a hierarchical taxonomy of terms, and a processor, which uses the taxonomy to perform effective mining of the database. Preferably, the system enables quantitative, content-based, textual analysis of a large number of documents in the database, in order to present relationships between two or more entries in the taxonomy.

Brief Summary Text (14):
Preferably, a user provides an input indicating terms of interest (some or all of which may be in the taxonomy), and the processor subsequently discovers relationships between terms in the user's input and terms in the taxonomy. Typically, relationships discovered during text mining comprise co-occurrences of two terms in a single document. Preferably, if the user "selects" one of the relationships generated by the text analysis, the system displays relevant portions of original documents in the database which are associated with the discovered relationship.

Brief Summary Text (15):
In some preferred embodiments of the present invention, terms in the term taxonomy ("taxonomy terms") can be edited by the user prior to text mining, and the taxonomy can be modified automatically by the processor and/or interactively with the user, responsive to results of the text mining. Typically, interactive editing of the term taxonomy responsive to results of the text mining yields improved results from a subsequent iteration of text mining, and these improved results may themselves be used to modify the taxonomy again. In this manner, the user may derive information of increased value from each iteration of text mining and term taxonomy modification.

Brief Summary Text (16):
In some preferred embodiments of the present invention, the taxonomy generally has a Directed Acyclic Graph (DAG) structure or a tree structure, and comprises groups of related terms (siblings) stored in the hierarchy one level below respective parent entries. For example, under a parent entry, "Countries," the taxonomy may contain as daughter entries the list of member nations of the United Nations. (The parent entry "Countries" may itself also be a member of a set of siblings in the taxonomy, under a "grandparent" entry, "Political entities.") Prior to text mining, in this example, the user may add the name of a new member nation, or delete the name of a country whose name has changed. Following text mining of the database, and utilizing results derived therefrom, the user may choose to further edit the term taxonomy (for instance, by adding a new country name or variation thereof).

Brief Summary Text (17):
In some preferred embodiments, the taxonomy has multiple levels, and a broad range of terms in each level, so that the user can narrow or broaden a query prior to an iteration of text mining, in order to optimize the results generated by the processor. For example, if the user would like to investigate President Clinton's foreign policy, she might enter an initial query specifying "Clinton" and all daughter entries of the node "Countries." To broaden the query, "Countries" could be replaced by "Political entities," so that a news article, containing the words "Berlin" and "Paris," but not "Germany" and "France," would also generate a positive response to the query. Alternatively, to narrow the query, the user could specify a taxonomy node "G7 countries," instead of "Countries." In general, a rich, multilevel taxonomy enables the user to enter queries with a desired level of specificity, and to thereby obtain information most relevant to her needs.

Brief Summary Text (18):
In a preferred embodiment, the processor prompts the user to refine the query prior to mining of the database's text, in order to optimize the results generated by the processor. For example, if the user enters a query including the words "Colombia" and "Venezuela," the processor preferably examines the taxonomy, determines that the two terms are daughter entries of a parent entry, "South American countries," and asks the user whether the two specified terms should be replaced by the names of all of the countries in South America listed in the taxonomy. Alternatively or additionally, the processor examines daughter entries of "Colombia" and "Venezuela," and asks the user whether some or all of the daughter entries (for instance, names of cities or politicians) should be added to the query.

Brief Summary Text (19):
In preferred embodiments of the present invention, text mining typically includes determining relationships among terms found in the database which relate to the user's query. Preferably, according to some preferred embodiments of the present invention, the processor subsequently uses these discovered relationships in order to suggest modifications to the taxonomy. For example, if the user's query includes the word "Venezuela" and a taxonomy node "Natural

resources," then text mining of the database may determine that the terms "Crude oil," "Coffee," "Sugarcane," and "Bananas" occur with high frequency in documents in the database having the word "Venezuela" and at least one daughter entry of "Natural resources". If, from this list, only "Sugarcane" is not a daughter entry of "Natural resources," then the processor preferably prompts the user to indicate whether "Sugarcane" should be added to the taxonomy as a daughter entry of "Natural resources". Should the user agree, then in processing a subsequent query including, for example, a taxonomy node "Ecological issues" and the taxonomy node "Natural resources", the processor will already "know" that sugarcane is a natural resource. In this manner, useful information derived by the text mining process is reported to the user, and is additionally used to improve the taxonomy in order to enhance the effectiveness of subsequent mining of the same or a different database.

Brief Summary Text (20):
Alternatively or additionally, the results of text mining may indicate to the user that a new node should be added to the taxonomy, or that an existing node should be supplemented in light of the generated results. For example, during text mining of a news database, the inventors entered a query including the term, "Ford Motor Corp.," and the taxonomy node, "Companies," so that the processor would generate a list of companies ranked by their frequency of co-occurrence with Ford. Most of the top 10 companies listed were car companies, and this might suggest to the user to create a new node, "Car companies," and to copy the appropriate companies into the new node.

Brief Summary Text (23):
In a similar manner, specific events can be extracted from a document's text. For example, "merger" is a relationship term known to link the names of two companies. If the word "merger" were found in the text of a document, the program would scan the "Companies" node in the taxonomy and report if two known company names are found in the vicinity of "merger," and are grammatically linked to "merger" according to predetermined rules.

Brief Summary Text (25):
There is therefore provided in accordance with a preferred embodiment of the present invention, a method for mining in a database including documents that include text. The method comprises providing a taxonomy of taxonomy terms, mining the documents responsive to the taxonomy to discover a relationship between two or more of the taxonomy terms, analyzing occurrences of the relationship over a plurality of the documents to extract information not specified by the taxonomy relation to the two, or more taxonomy terms, and persenting the information relating to the two or more taxonomy terms to a user.

Brief Summary Text (26):
In a preferred embodiment, analyzing occurrences of the relationship comprises identifying, responsive to the taxonomy, one of: a fact and an event, inherent in the text of one of the documents. The fact or event may be identified by the proximity, in one of the documents, of the two or more taxonomy terms to a predetermined relationship term.

Brief Summary Text (27):
Preferably, the taxonomy comprises nodes and one of the nodes is a parent entry of the two or more taxonomy terms. Alternatively or additionally, the taxonomy comprises a hierarchy of nodes, wherein a first node is a parent entry of at least one of the two or more taxonomy terms, wherein a second node is also a parent entry of at least one of the two or more taxonomy terms, and wherein the relationship comprises a relationship between the second node and the first node.

Brief Summary Text (28):
Preferably, analyzing comprises analyzing, over a plurality of the documents, co-occurrences of substantially every one of the two or more taxonomy terms with substantially every other one of the two or more taxonomy terms, to determine relationships among the two or more taxonomy terms; and presenting the information comprises displaying at least some of the two or more taxonomy terms and displaying an output indicative to the number of the co-occurrences of substantially each of the at least some terms with substantially every other one of the at least some terms.

Brief Summary Text (30):
Preferably, mining includes extracting a set of one or more document-labeling terms from one of the documents, wherein the set includes the two or more taxonomy terms. Extracting the set of document-labeling terms from the document may comprise determining the grammatical structure of a sentence in the document's text and identifying a group of one or more words in the sentence as a document-labeling term responsive to the grammatical structure. Preferably, extracting the set of document-labeling terms from the document may comprises: examining the document, identifying a candidate term in the document, comparing a frequency of occurrence of the candidate term in the document with frequencies of occurrence of the candidate term in other documents in the database to determine differences in the respective frequencies of occurrence, and inserting the candidate term into the set of document-labeling terms corresponding to the document responsive to the comparison.

Brief Summary Text (31):
In one preferred embodiment, discovering the relationship comprises finding in at least one of the documents a co-occurrence of at least some portion of the two or more taxonomy terms. The two or more taxonomy terms may comprise a cluster of taxonomy terms, wherein the cluster is characterized by the property that terms in the cluster generally have a higher frequency of co-occurrence with respect to each other than their frequency of co-occurrence with respect to terms not in the cluster. Also, discovering the relationship may comprise assigning a weight to the co-occurrence responsive to a distance between a first term and a second term of the two or more taxonomy terms, and wherein analyzing occurrences of the relationship comprises analyzing the relationship responsive to the weight. For example, a first term and a second term of the two or more taxonomy terms may co-occur in a document when a distance between the first and second terms is less than a predetermined distance. The predetermined distance may be, for instance, approximately one paragraph or approximately one sentence.

Brief Summary Text (32):
In another embodiment, discovering the relationship comprises discovering a plurality of relationships in two or more of the documents where analyzing comprises: analyzing the relationships in a first set of the two or more documents, in order to determine a first relationship between the two or more taxonomy terms; analyzing the relationships in a second set of the two or more documents, in order to determine a second relationship between the two or more taxonomy terms; and comparing the first and second relationships. The first set of documents may comprise documents from a first time period, and the second set of documents may comprise documents from a second time period.

Brief Summary Text (33):
In a preferred embodiment, at least one of the two or more taxonomy terms is selected by the user, and, in one embodiment, each of the two or more taxonomy terms is selected by the user. Also, presenting the information may comprise displaying a graph comprising a plurality of points, each point representing one of the two or more taxonomy terms, and one or more lines, each line connecting two of the points and indicating a quantitative relationship between the terms represented by said two points. For example, the thickness of each line in the graph, a number displayed near each line in the graph, or the color of each line in the graph may indicate the quantitative relationship. The quantitative relationship indicated by a line in the graph is preferably a co-occurrence frequency of the terms represented by the two points connected by that line.

Brief Summary Text (34):
There is further provided, in accordance with a preferred embodiment of the present invention, a method for mining in a database including documents, the documents including text. The method comprises providing a taxonomy of taxonomy terms, mining the documents responsive to the taxonomy to discover a relationship between a set of one or more selected words and at least one of the taxonomy terms, and modifying the taxonomy responsive to the discovered relationship. The taxonomy may comprise a hierarchy of nodes, wherein the at least one taxonomy term comprises two or more related taxonomy terms, one of the nodes is a parent entry of the two or more taxonomy terms, and wherein modifying comprises assigning one of the selected words to be a sibling of the two or more taxonomy terms responsive to the discovered relationship.

Brief Summary Text (35):
The present invention additionally provides a method for mining in a database including documents, the documents including text, where the method comprises providing a taxonomy of taxonomy terms, mining the documents at a taxonomy term level to provide mining results indicative of a relationship between a plurality of terms including at least one taxonomy term, wherein at least one of the terms is specified by a user, and prompting the user to modify the taxonomy based on the mining results. The method preferably further comprises performing a statistical analysis on the mining results to determine a potential modification of the taxonomy and prompting the user on whether or not to carry out the potential modification of the taxonomy. At least one of the plurality of terms may be received in a query entered by the user. In one embodiment, at least one of the plurality of terms is a taxonomy term specified by the user.

Brief Summary Text (36):
The present invention yet further provides a method for mining in a database including documents, the documents including text, with the method comprising providing a taxonomy of taxonomy terms; receiving a query from a user, the query specifying at least one term of interest; mining the documents at a taxonomy term level to provide a first set of mining results indicative of a relationship between the at least one term of interest and at least one of the taxonomy terms; modifying the taxonomy based, at least in part, on the first set of mining results; and mining the documents at a modified taxonomy term level to provide a second set of mining results indicative of the relationship. The method may further comprise, in response to the query, displaying to the user a portion of the taxonomy relevant to the at least one term of interest, and enabling the user to revise the query by including in the query at least one of the taxonomy terms in the displayed portion.

Brief Summary Text (37):
The present invention still further provides a method for mining in a database including documents, the documents including text, the method comprising providing a taxonomy of taxonomy terms; receiving an initial query from a user, the query specifying at least one term of interest; displaying to the user a portion of the taxonomy relevant to the at least one term of interest; receiving an indication from the user to revise the query by including in the query at least one taxonomy term in the displayed portion; and mining the documents at a taxonomy term level based on the revised query to provide mining results indicative of a relationship between the at least one taxonomy term and one or more other terms.

Brief Summary Text (38):
The present invention also provides a method for mining in a database including documents, the documents including text, the method comprising providing a taxonomy of taxonomy terms; mining the documents at a taxonomy term level to provide mining results indicative of a relationship between a plurality of terms including at least one taxonomy term, wherein at least one of the terms is specified by a user; and presenting the mining results to the user by displaying a graph comprising a plurality of points, each point representing one of the plurality of terms, and one or more lines, each line connecting two of the points and indicating a quantitative relationship between the terms represented by said two points.

Drawing Description Text (5):
FIG. 3 is a flow chart schematically illustrating a method for editing a taxonomy, for use in executing the method of FIG. 1B, in accordance with a preferred embodiment of the present invention;

Drawing Description Text (6):
FIG. 4 is a flow chart schematically illustrating details of the taxonomy editing method of FIG. 3, in accordance with a preferred embodiment of the present invention;

Detailed Description Text (2):
FIG. 1A is a schematic diagram illustrating a system 20, comprising a controller 24 which performs text mining in a database 36, in accordance with a preferred embodiment of the present invention. Examples of text mining operations, as provided by the present invention, are

described hereinbelow with reference to FIG. 2. Controller 24 preferably comprises a text mining processor 26, a taxonomy storage unit 28, a monitor 32, and an input device 38. Database 36 is typically coupled to controller 24 by a network 34, as shown in FIG. 1A, but may alternatively be located within controller 24.

Detailed Description Text (3):
Preferably, taxonomy storage unit 28 comprises a hard disk or other mass storage device, upon which is stored a hierarchical term taxonomy. Typically, the taxonomy has a Directed Acyclic Graph (DAG) structure or a tree structure, such as that shown in Table I hereinbelow, in which are stored a large number of terms, some of which are preferably related, similar or identical to terms in a query entered by a user 22 of system 20. For example, under a parent entry, "Companies," there may be a series of daughter entries, including "Ford Motor Corp.," "Merck Corp.," and "Microsoft Corp." (Typically, broad nodes such as "Companies" or "Names" have thousands of daughter entries.) Under the entry "Microsoft Corp.," there may be additional daughter entries, "Management," "Employees," "Products," "Development facilities," and "Major shareholders. " As described hereinbelow, the taxonomy is preferably used prior to, during, and/or following a text mining iteration, in order to add power and effectiveness to controller 24's implementation of user 22's query. Alternatively, other data structures known in the art of data storage and retrieval are used to store the taxonomy.

Detailed Description Text (4):
Text mining processor 26 preferably comprises a central processing unit (CPU), which executes one or more programs to process user 22's query, carry out text mining responsive thereto, and, typically, semi-automatically modify the taxonomy responsive to the results of the text mining, as described hereinbelow.

Detailed Description Text (5):
Database 36 typically comprises documents, each comprising text. For example, some of the documents may be news stories, articles in an encyclopedia, books, lists, or other collections of text stored in a digital form. Alternatively or additionally, the one or more documents in database 36 may comprise respective sentences, paragraphs, or sections, from a single file containing text. It will be understood by one skilled in the art that although preferred embodiments of the present invention are described with respect to text stored in each document, the documents may additionally contain video, audio, or other data forms which typically are not processed during application of the present invention. Alternatively, means are employed to extract textual data from audio or graphical components of a document, for example by applying speech-to-text algorithms or optical character recognition algorithms.

Detailed Description Text (7):
FIG. 1B is a flow chart schematically illustrating a method 46 for text mining in database 36, in accordance with a preferred embodiment of the present invention. Method 46, according to the present invention, preferably comprises two main steps: "mine text at the term level" 40 and "edit taxonomy" 30. Mining step 40 preferably receives the query from user 22, examines and analyzes the taxonomy, and, optionally, assists the user in refining the query in light of the analysis, using methods described hereinbelow. Subsequently, mining step 40 performs the text mining, as described hereinbelow, and displays the results of the text mining obtained thereby. In taxonomy editing step 30, processor 26 preferably analyzes the results of the text mining and suggests modifications to the taxonomy responsive to this analysis.

Detailed Description Text (8):
For example (using the sample taxonomy shown in Table I), if the user's initial query includes the term "Microsoft Corp.," then, prior to text mining, processor 26 preferably displays the relevant section of the taxonomy, and enables the user to indicate whether the term "Microsoft Corp.," in particular, should be used in implementing the query, or whether, in addition, documents including daughter entries of one or more of "Management," "Employees," "Products," "Development facilities," and "Major shareholders,"--but not the term "Microsoft Corp. " itself--should be retrieved during searching of the database for relevant documents. In this manner, depending on the user's response, an article citing only "Bill Gates" or "MS-DOS" may be labeled as relevant to the user's query.

Detailed Description Text (9):
Following analysis of the results of the text mining, processor 26 may ask user 22 whether other terms, which were found during text mining to be highly correlated with some or each of the daughter entries of "Products," should be added to the taxonomy as siblings of existing daughter entries "MS-DOS," "Microsoft Office," "Microsoft Works," "Microsoft Publisher," and "Microsoft Word. " In this example, the processor might suggest adding the terms, "Microsoft Excel," "Microsoft Photo Editor," "Microsoft Outlook," and "Desktop computer. " User 22 may choose to add none, some, or all of the suggested terms at the proposed location in the taxonomy, and/or at one or more other nodes. Typically, as in this example, some of the suggestions are appropriate, while others (i.e., "Desktop computer") are highly correlated with the siblings in the suggested node, but nevertheless would not generally be chosen by the user for insertion into the taxonomy.

Detailed Description Text (11):
Preferably, input block 82 comprises the inputs to the text mining process, including database 36, the taxonomy, the user's query, and a "term generator" 80 for the set of documents. Term generator 80 is typically implemented using an algorithm similar to that described in the above-mentioned paper, "Text Mining at the Term Level," and is described hereinbelow with reference to FIG. 7. Terms are typically defined based on grammatical rules, such as "Noun-noun" (e.g., "Stock trader"), "Noun-preposition-noun" (e.g., "King of Jordan"), etc. Preferably, term generator 80 extracts from each document in database 36 a set of terms which generally represent the contents of the respective document. For example, term generator 80 analyzed an article in the Reuters Financial News Database, and identified terms like "Net income," "Bank," "Earnings," "Canada," "Mutual Fund," and "National Bank of Canada" as being particularly relevant to the content of the article, while it rejected terms like "Jump," "Season," and "Group," for being not particularly relevant to the article's content.

Detailed Description Text (12):
Reference is now made to FIGS. 2, 3 and 4. FIG. 3 is a flow chart schematically illustrating details of taxonomy editing step 30, in accordance with a preferred embodiment of the present invention. FIG. 4 is a flow chart schematically illustrating details of steps 50 and 60 of editing step 30, in accordance with a preferred embodiment of the present invention. A variety of useful text mining functions are provided by the present invention, as described hereinbelow. In applying these functions, the present invention provides the user with the ability to essentially continuously and interactively revise the taxonomy responsive to results of text mining shown by display block 44, and to use the revised taxonomy to refine the query, in order to enhance the user's ability to derive useful information from system 20.

Detailed Description Text (13):
As described hereinabove with reference to FIG. 1B, and as shown in FIG. 2, the taxonomy is preferably used to optimize the user's query, in order to add power and effectiveness to the text mining process. For example, if the user is interested in discovering computer companies having significant dealings in the United States and France, then, in general, the query should include the nodes "Computer companies," "United States," and "France. " Processor 26 preferably displays relevant sections of the taxonomy, and allows the user to indicate whether all or some of the following terms should be added to the query as optional replacements for the corresponding terms in the original query: {"Microsoft", "IBM", "Apple", "Compaq"}, {"Washington", "America", "United States of America", "USA", "American"}, and {"Paris", "French", "European Community"}. Additionally, parent entries, daughter entries, granddaughter entries, etc., of any chosen term can be specified by the user to be included in the query. Should each of the cited terms be chosen to be added to the query, then a news article, whose list of terms created by generator 80 includes {"Microsoft", "America", "France"} would be available for analysis by text mining, even though it does not contain even one word from the original query. Furthermore, after a first iteration of text mining, processor 26 may find in step 50 (FIGS. 3 and 4) that each of the following terms occurs with high frequency in documents matching the revised query, compared to the terms' frequencies in all documents in database 36: {"White House", "US", "Trade controls", "Copyright protection", "Dell", "Bill Clinton", "Jacques Chirac", "European Parliament"}.

Detailed Description Text (14):

In and of themselves, these results may provide user 22 with information that she previously had not had, and which might not have been immediately available by simply applying a search engine to the same, hypothetical, news database (even with the revised query) . To attain even greater power from the text mining process, user 22 is prompted in step 60 to indicate whether any of the "Result" terms should be incorporated into the taxonomy. Preferably, insertions into and editing of the taxonomy are performed in step 70 using user interface 38 (FIG. 1A), which typically comprises a keyboard and mouse. Most preferably, user 22 modifies the taxonomy using well-known editing tools, such as "drag-and-drop" editing, "double-clicking" upon desired entries, use of dialog boxes, etc. In the present example, user 22 might drag "White House" and "US" into the taxonomy so that these terms become siblings of "United States of America. " Similarly, "Jacques Chirac" may be made a daughter entry of "France. "

Detailed Description Text (15):
Typically, changes to the taxonomy are immediately used in a subsequent text mining iteration. Thus, the inclusion of the terms "Jacques Chirac," "White House," and "US" may generate a larger set of useful, relevant, documents, upon which the analysis functions provided by the present invention can be performed.

Detailed Description Text (16):
Alternatively or additionally, user 22 may choose to edit the taxonomy substantially independently of the results of the text mining process. This "manual" editing of the taxonomy typically comprises copying an external taxonomy into the current taxonomy. External taxonomies generally comprise lists, such as the names of bones in the human body, the countries and capitals of South America, the names of the planets, etc. Additionally, user 22 is preferably enabled to move or copy sub-trees of the taxonomy, or to add, delete, or edit individual entries in the taxonomy.

Detailed Description Text (17):
In a preferred operational mode, an interactive browser 92 of terms processor 90 (FIG. 2) enables the user to enter a query, to see the results of a first iteration of text mining responsive to the query, and to enter a new query based on the knowledge obtained during the first text mining iteration. In an experimental application of the interactive browser, the inventors investigated co-occurrences of daughter entries of two nodes in the taxonomy, "Business alliance topics" and "Companies," as reflected in the approximately 52,000 articles contained in the Reuters Financial News Database from 1995 and 1996. Prior to input of the query, term generator 80 in input block 82 processed the 52,000 articles, which contained an average of 864 words per article, and produced thereby an average of 45 terms per article. Prior to text mining, the taxonomy contained a set of terms, such as "Joint venture," "Strategic alliance," and "Merger," under the parent entry, .sup.1 "Business alliance topics," and an extensive list of company names under the parent entry, "Companies. " With input block 82 thus defined (term generation, database, taxonomy, and user query), interactive browser 92 found 569 relevant references in the Reuters database, generated a list of article titles, and highlighted terms which matched the user's query, e.g., {"Apple Computer", "Sun Microsystems," "Merger talk"}, {"Lockheed Corp. ", "Martin Marietta Corp. ", "Merger"}, and {"Chevron Corp. ", "Mobil Corp. ", "Joint venture"}.

Detailed Description Text (18):
Additionally, while using interactive browser 92, user 22 can employ a distribution browser 94 of terms processor 90, in order to cause a statistical analysis to be performed on the text mining results. For example, in the above-cited experiment, the distribution of all daughter entries of "Business alliance topics" was calculated, revealing that "Joint venture" had the highest frequency of occurrence (relative to its siblings) in the 569 references. A "Company" distribution was subsequently calculated on those articles containing the term "Joint venture," showing that IBM, General Motors and MCI were the companies that co-occurred most frequently with "Joint venture. " It was additionally determined that Sprint Corp. and News Corp. were the two other companies which occurred most frequently in those articles in which MCI co-occurred with "Joint venture. " Lastly, utilizing an additional node in the taxonomy, "People," a further iteration of text mining found that "Jeffrey Kagan," "Anthea Disney," and "Bill Vogel" were the three people most frequently cited in "Joint venture" articles which referred to MCI and News Corp. This example demonstrates the added power that interactive analysis of the

results of text mining can yield, especially when used in conjunction with a rich taxonomy.

Detailed Description Text (19):
In some preferred embodiments of the present invention, interactive browser 92 discovers in one or more documents a specific fact or event using terms that are in the taxonomy, such as those described hereinabove, and a set of "relationship terms," which are not necessarily in the taxonomy. Relationship terms comprise words that often are an integral part of a factual statement or implication inherent in a text, and that additionally often indicate a relationship between two terms in the same sentence. Sample relationship terms, and the types of terms which they link, include: "President" (Name⇄Country or Company), "Senator" (Name⇄State), "Capital" (Place⇄Country), "Longest" (Noun⇄Noun), "Hired" (Company⇄Name), and "Joint venture" (Company⇄Company).

Detailed Description Text (20):
For example, the fact that "Bill Clinton" is the "President" of "the United States of America" can be deduced from a sentence in the middle of a document in the database, "US President Bill Clinton addressed a trade meeting on Thursday. " Preferably, a list of synonyms is already known to the processor, for example, "US"="the United States of America. " Even in cases where a relationship term and only one of the corresponding taxonomy terms is present, e.g., "President Clinton appeared invigorated after a successful ten-day trip," a useful fact can nevertheless often be gleaned from the sentence ("Clinton" is the "President").

Detailed Description Text (21):
In a similar manner, specific events can be extracted from a document's text. For example, "merger " is a relationship term known to link the names of two companies. If the word "merger" were found in the text of a document, a program implementing this embodiment would scan the "Companies" node in the taxonomy and report if two known company names are found in the vicinity of "merger," and are grammatically linked to "merger" according to predetermined rules.

Detailed Description Text (23):
Often, the group profile analyzer, used independently or with the other text mining tools, quickly gives the user an overall sense of the contents of the database, with respect to a node in the taxonomy (e.g., G7 countries, local hospitals, Ivy League universities), without the need for reading even a single article. Subsequently, the user can optionally see the actual text of one or more articles of interest (for example, those listing the names of each of the G7 countries and the word "Livestock").

Detailed Description Text (24):
In some applications, user 22 may use group profile analyzer 96 in order to suggest modifications to the taxonomy. For example, if it is found that fifteen out of twenty siblings are relatively close to the overall group profile, while five of the siblings are relatively far, the user might consider examining the five to see whether they should be split off to form a new node (e.g., the five siblings might be supercomputers, listed as siblings of fifteen personal computers).

Detailed Description Text (25):
Other investigations, performed using a "profile comparison" block 98 of terms processor 90, include comparisons of two different groups, for example G7 countries and the Arab League, in order to determine topics which are highly associated in the database with one or the other group (e.g., "Corporate bonds" or "Crude oil"). In this manner, user 22 is enabled to quickly determine similarities and/or differences between groups, without reading a single article in the database. An additional investigation, using a query including the taxonomy nodes "Caffeine-based drinks," "Country groups," and "Countries," revealed that all countries that are highly correlated with caffeine-based drinks belong either to the OAU (Organization of African Unity) or to the OAS (Organization of American States).

Detailed Description Text (26):
A "high-correlation pairs" block 102 of terms processor 90 preferably automatically locates, and presents to user 22, pairs of terms in database 36 (or in a predetermined cross-section

thereof) that are highly correlated with each other. The inventors have found, in some applications, that block 102 is likely to present the user with important relationships within the database or the cross-section thereof. For example, in the cross-section of a news database having only articles including terms from a node, "Political terms," such important relationships may include a politician's name and the issues with which he is most associated. It is emphasized that the text mining tools of terms processor 90, in combination with interactive taxonomy editing, as provided by the present invention, have shown the ability (in the inventors' opinion) to extract knowledge, or useful content, from the database, and not simply to generate a list of articles that match terms in the user's query.

Detailed Description Text (31):
Typically, a keyword graph is defined with respect to a "context," usually an additional term or a node in the taxonomy. For example, FIG. 5 shows the results of a query, including the node "Countries" and the term "Crude oil," where the keyword graph represents countries as vertices, and the thickness of the edge connecting two vertices reflects the number of articles in the Reuters database containing both country names.. An adjustable threshold of six articles containing a given two countries was set as a minimum requirement for drawing an edge therebetween, in order to increase the clarity of the graph, and to allow the most significant data to be clearly visible. It should be noted that a substantial quantity of meaningful data is displayed in FIG. 5, and that user 22 does not have to read any articles in the database in order to see significant facts about the relationships among the countries shown.

Detailed Description Text (32):
In another embodiment (not shown), the number of articles represented by an edge is alternatively or additionally written next to the edge, in order to enable a quantitative comparison of the number of articles represented by each edge. In still another embodiment (not shown), the color or darkness of each edge is related to the co-occurrence frequency of the respective vertices.

Detailed Description Text (34):
A "trend graphs" block 110 of terms processor 90 (FIG. 2) is preferably used in conjunction with one or more of the other blocks of processor 90, in order to display significant changes in a given text mining result from a first time point to a second time point, or to construct a graph, which displays a text mining output parameter (e.g., the number of co-occurrences of "Microsoft" and "Justice Department") with respect to a time axis. Using block 110, is enabled to quickly discover the time course of the news coverage of a given issue. In an investigation using the Reuters database covering the early 1980's, the inventors asked, from those articles which mention Libya, what is the percentage that also mention Chad. It was determined using block 110 that the percentage went from 0% in a first three-month period to 35% in the following three-month period. Double-clicking on the line representing the second three-month period showed news articles dealing exclusively with fighting between Libya and Chad.

Detailed Description Text (36):
FIG. 6 is a graph illustrating another output of the text mining algorithm of FIG. 2, in a format which is particularly suitable for displaying a large number of terms, in accordance with a preferred embodiment of the present invention. In this example, user 22's query included the taxonomy nodes "Companies" and "Computer technologies," and was used to analyze a news database. Notably, the user specified only the two taxonomy nodes, the database, and the output format. Responsive thereto, processor 26 reviewed all of the terms (generated by term generator 80) associated with each of the articles in the database, and recorded instances in which one of the companies co-occurred in the same sentence with one of the specified computer technologies. Subsequently, processor 26 tabulated the results, and displayed, as shown in FIG. 6, only those companies and technologies whose co-occurrences in the database surpassed a user-set threshold. The thickness of the lines connecting respective companies and technologies represents, as in FIG. 5, the respective number of co-occurrences.

Detailed Description Text (39):
Preferably, the user can select a "co-occurrence proximity threshold," i.e., the maximum distance between two terms in a query such that processor 26 should record a correlation between the two terms. Typically, the co-occurrence proximity threshold is specified as a

certain number of words, sentences or paragraphs, or, simply, as co-occurrence within the same article. In general, increasing the proximity threshold increases the number of articles deemed relevant during text mining but tends to decrease the specificity, or "value," of the identified co-occurrence. For example, the user typically would not want processor 26 to report an association between the terms "Iraq" and "Tobacco," if the two terms' closest proximity in an article is several paragraphs. The inventors have found that a co-occurrence proximity threshold of approximately one sentence is generally optimal for many applications of the present invention.

Detailed Description Text (40):
Alternatively or additionally, a "weight" is assigned to a co-occurrence of two terms in a document responsive to the proximity of the two terms, so that, for example, a consistent trend of co-occurrences of the terms "Iraq" and "Tobacco" in documents in the database will be reported, even if in all of the documents the terms are separated by a relatively large distance.

Detailed Description Text (42):
As described hereinabove, the user is enabled to set a threshold number of co-occurrences, below which a line will not be drawn connecting two terms. This typically serves two functions. First, it reduces the effect of "coincidental" co-occurrences, e.g., a correlation between "London" and "Jacques Chirac," based on the sentence, "While flying to London, President Clinton telephoned President Jacques Chirac to discuss trade issues. " Second, control of the threshold gives the user the ability to see a very large number of related terms in one graph, or, alternatively, to focus on the several co-occurrences which appear with the highest frequency in the database. The inventors have found that, generally speaking, a threshold value of between about three and five co-occurrences tends to significantly reduce the appearance of coincidental co-occurrences.

Detailed Description Text (49):
The test.drf file contains a representation of data in a document. The test.tax file contains a relevant taxonomy. The test.dre file is for internal usage by the software. When installing the software, it is recommended that all three files be located in the same directory as NKDT.exe.

Detailed Description Text (50):
In the microfiche appendix, all of the above files have been compressed using the ZIP compression program. The appendix contains a self-extractable ZIP file, which, when executed, provides all of the above files. To extract the file, copy the data provided in the microfiche into an appropriate computer, and run the application using the operating system "run" function.

Current US Original Classification (1):
707/6

Current US Cross Reference Classification (1):
707/10

CLAIMS:

1. A method for mining in a database including documents, the documents including text, the method comprising: providing a taxonomy of taxonomy terms; mining the documents responsive to the taxonomy to discover a relationship between two or more of the taxonomy terms; analyzing, over a plurality of the documents, co-occurrences of substantially every one of the two or more taxonomy terms with substantially every other one of the two or more taxonomy terms, to determine relationships among the two or more taxonomy terms to extract information not specified by the taxonomy relating to the two or more taxonomy terms; and presenting the information relating to the two or more taxonomy terms to a user by displaying at least some of the two or more taxonomy terms and displaying an output indicative of the number of the co-occurrences of substantially each of the at least some terms with substantially every other one of the at least some terms.

2. A method according to claim 1, wherein analyzing occurrences of the relationship comprises identifying, responsive to the taxonomy, one of: a fact and an event, inherent in the text of one of the documents.

3. A method according to claim 2 wherein the fact or event is identified by the proximity, in one of the documents, of the two or more taxonomy terms to a predetermined relationship term.

4. A method according to claim 1, wherein the taxonomy comprises nodes and one of the nodes is a parent entry of the two or more taxonomy terms.

5. A method according to claim 1, wherein the taxonomy comprises a hierarchy of nodes, wherein a first node is a parent entry of at least one of the two or more taxonomy terms, wherein a second node is also a parent entry of at least one of the two or more taxonomy terms, and wherein the relationship comprises a relationship between the second node and the first node.

7. A method according to claim 1, wherein mining comprises extracting a set of one or more document-labeling terms from one of the documents, wherein the set includes the two or more taxonomy terms.

8. A method according to claim 7, wherein extracting the set of document-labeling terms from the document comprises: determining the grammatical structure of a sentence in the document's text; and identifying a group of one or more words in the sentence as a document-labeling term responsive to the grammatical structure.

9. A method according to claim 7, wherein extracting the set of document-labeling terms from the document comprises: examining the document; identifying a candidate term in the document; comparing a frequency of occurrence of the candidate term in the document with frequencies of occurrence of the candidate term in other documents in the database to determine differences in the respective frequencies of occurrence; and inserting the candidate term into the set of document-labeling terms corresponding to the document responsive to the comparison.

10. A method according to claim 1, wherein discovering the relationship comprises finding in at least one of the documents a co-occurrence of at least some portion of the two or more taxonomy terms.

11. A method according to claim 10, wherein the two or more taxonomy terms comprise a cluster of taxonomy terms, wherein the cluster is characterized by the property that terms in the cluster generally have a higher frequency of co-occurrence with respect to each other than their frequency of co-occurrence with respect to terms not in the cluster.

12. A method according to claim 10, wherein discovering the relationship comprises assigning a weight to the co-occurrence responsive to a distance between a first term and a second term of the two or more taxonomy terms, and wherein analyzing occurrences of the relationship comprises analyzing the relationship responsive to the weight.

13. A method according to claim 10, wherein a first term and a second term of the two or more taxonomy terms co-occur in a document when a distance between the first and second terms is less than a predetermined distance.

16. A method according to claim 1, wherein discovering the relationship comprises discovering a plurality of relationships in two or more of the documents, and wherein analyzing comprises: analyzing the relationships in a first set of the two or more documents, in order to determine a first relationship between the two or more taxonomy terms; analyzing the relationships in a second set of the two or more documents, in order to determine a second relationship between the two or more taxonomy terms; and comparing the first and second relationships.

18. A method according to claim 1 wherein at least one of the two or more taxonomy terms is selected by the user.

19. A method according to claim 1 wherein each of the two or more taxonomy terms is selected by

the user.

20. A method according to claim 1 wherein presenting the information comprises displaying a graph comprising a plurality of points, each point representing one of the two or more taxonomy terms, and one or more lines, each line connecting two of the points and indicating a quantitative relationship between the terms represented by said two points.

24. A method according to claim 20 wherein the quantitative relationship indicated by a line in the graph is a co-occurrence frequency of the terms represented by the two points connected by that line.

25. A method for mining in a database including documents, the documents including text, the method comprising: providing a taxonomy of terms, the taxonomy comprising a hierarchy of nodes, wherein at least one taxonomy term comprises two or more related taxonomy terms, and one of the nodes is a parent entry of the two or more taxonomy terms, mining the documents responsive to the taxonomy to discover a relationship between a set of one or more selected words and at least one of the taxonomy terms; and modifying the taxonomy responsive to the discovered relationship by prompting a user to assign one or more of the selected words to be sibling in the taxonomy of the two or more taxonomy terms.

26. A method according to claim 25, wherein the relationship comprises a co-occurrence of one of the set of selected words and the at least one taxonomy term.

27. A computer program product for mining in a database including documents, the documents including text, the program having computer-readable program instructions embodied therein, which instructions cause a computer to: provide a taxonomy of taxonomy terms; mine the documents responsive to the taxonomy to discover a relationship between two or more of the taxonomy terms; analyze, over a plurality of the documents, co-occurrences of substantially every one of the two or more taxonomy terms with substantially every other one of the two or more taxonomy terms, to determine relationships among the two or more taxonomy terms and to extract information not specified by the taxonomy relating to the two or more taxonomy terms; and display at least some of the two or more taxonomy terms and display an output indicative of the number of the co-occurrences of substantially each of the at least some terms with substantially every other one of the at least some terms.

28. A product according to claim 27, wherein analyzing occurrences of the relationship comprises identifying, responsive to the taxonomy, one of: a fact and an event, inherent in the text of one of the documents.

29. A product according to claim 27, wherein the taxonomy comprises nodes and one of the nodes is a parent entry of the two or more taxonomy terms.

30. A product according to claim 27, wherein the taxonomy comprises a hierarchy of nodes, a first node is a parent entry of at least one of the two or more taxonomy terms, wherein a second node is also a parent entry of at least one of the two or more taxonomy terms and wherein the relationship comprises a relationship between the second node and the first node.

31. A product according to claim 27, wherein: analyzing comprises analyzing, over a plurality of the documents, co-occurrences of substantially every one of the two or more taxonomy terms with substantially every other one of the two or more taxonomy terms, to determine relationships among the two or more taxonomy terms; and presenting the information comprises displaying at least some of the two or more taxonomy terms and displaying an output indicative to the number of the co-occurrences of substantially each of the at least some terms with substantially every other one of the at least some terms.

33. A product according to claim 27, wherein mining comprises extracting a set of one or more document-labeling terms from one of the documents, wherein the set includes the two or more taxonomy terms.

34. A product according to claim 33, wherein extracting the set of document-labeling terms from

the document comprises: determining the grammatical structure of a sentence in the document's text; and identifying a group of one or more words in the sentence as a document-labeling term responsive to the grammatical structure.

35. A product according to claim 33, wherein extracting the set of document-labeling terms from the document comprises: examining the document; identifying a candidate term in the document; comparing a frequency of occurrence of the candidate term in the document with frequencies of occurrence of the candidate term in other documents in the database to determine differences in the respective frequencies of occurrence; and inserting the candidate term into the set of document-labeling terms corresponding to the document responsive to the comparison.

36. A product according to claim 27, wherein discovering the relationship comprises finding in at least one of the documents a co-occurrence of at least some portion of the two or more taxonomy portions.

37. A product according to claim 36, wherein the two or more taxonomy terms comprise a cluster of taxonomy terms, wherein the cluster is characterized by the property that terms in the cluster generally have a higher frequency of co-occurrence with respect to each other than their frequency of co-occurrence with respect to terms not in the cluster.

38. A product according to claim 36, wherein discovering the relationship comprises assigning a weight to the co-occurrence responsive to a distance between a first term and a second term of the two or more taxonomy terms, and wherein analyzing occurrences of the relationship comprises analyzing the relationship responsive to the weight.

39. A product according to claim 36, wherein a first term and a second term of the two or more taxonomy terms co-occur in a document when a distance between the first and second terms is less than a predetermined distance.

42. A product according to claim 27, wherein discovering the relationship comprises discovering a plurality of relationships in two or more of the documents, and wherein analyzing comprises: analyzing the relationships in a first set of the two or more documents, in order to determine a first relationship between the two or more taxonomy terms; analyzing the relationships in a second set of the two or more documents, in order to determine a second relationship between the two or more taxonomy terms; and comparing the first and second relationships.

44. A method for mining in a database including documents, the documents including text, the method comprising: providing a taxonomy of taxonomy terms; mining the documents at a taxonomy term level to provide mining results indicative of a relationship between a plurality of terms including at least one taxonomy term, wherein at least one of the terms is specified by a user; and displaying a graph comprising a plurality of points, each point representing one of the plurality of terms, and one or more lines, each line connecting two of the points and indicating a quantitative relationship between the terms represented by said two points; and prompting the user to modify the taxonomy based on the mining results.

45. A method according to claim 44 further comprising performing a statistical analysis on the mining results to determine a potential modification of the taxonomy.

46. A method according to claim 45 wherein prompting the user to modify the taxonomy comprises prompting the user on whether or not to carry out the potential modification of the taxonomy.

48. A method according to claim 47, wherein the taxonomy comprises a hierarchy of nodes, one of the nodes being a parent entry of the at least one taxonomy term and one or more additional taxonomy terms related thereto, and wherein discovering the relationship comprises determining that another of the plurality of taxonomy terms co-occurs with a cohort of the related terms, wherein the cohort is of at least a minimum size.

49. A method according to claim 48 wherein each of the plurality of terms is a taxonomy term.

50. A method according to claim 44 wherein at least one of the plurality of terms is a taxonomy

term specified by the user.

51. A method according to claim 44 wherein at least two of the plurality of terms are taxonomy terms specified by the user.

52. A method according to claim 44, wherein discovering the relationship comprises finding, in at least one of the documents, a co-occurrence of the at least one taxonomy term with another of the plurality of taxonomy terms.

53. A method according to claim 52, wherein the at least one taxonomy term comprises a cluster of taxonomy terms, wherein the cluster is characterized by the property that taxonomy terms in the cluster generally have a higher frequency of co-occurrence with respect to each other than their frequency of co-occurrence with respect to taxonomy terms not in the cluster.

54. A method according to claim 52, wherein discovering the relationship comprises assigning a weight to the co-occurrence responsive to a distance between the terms, and wherein analyzing occurrences of the relationship comprises analyzing the relationship responsive to the weight.

55. A method according to claim 52, wherein discovering the relationship comprises finding, in at least one of the documents, a co-occurrence of the at least one taxonomy term with another of the plurality of taxonomy terms where a distance between the terms term is less than a predetermined distance.

56. A method for mining in a database including documents, the documents including text the method comprising: providing a taxonomy of taxonomy terms; mining the documents at a taxonomy term level to provide mining results indicative of a relationship between a plurality of terms including at least one taxonomy term, wherein at least one of the terms is specified by a user; and presenting the mining results to the user by displaying a graph comprising a plurality of points, each point representing one of the plurality of terms, and one or more lines, each line connecting two of the points and indicating a quantitative relationship between the terms represented by said two points by displaying a number near each line.

58. A method according to claim 56 wherein the quantitative relationship indicated by a line in the graph is a co-occurrence frequency of the terms represented by the two points connected by that line.

Previous Doc Next Doc Go to Doc#

L19: Entry 1 of 9 File: USPT May 11, 2004

DOCUMENT-IDENTIFIER: US 6735592 B1
TITLE: System, method, and computer program product for a network-based content exchange system

Brief Summary Text (5):
For more information regarding such research, additional reference may be made to the following documents: Marti Hearst. 1992. Direction-Based Text Interpretation as an Information Access Refinement. In [Jacobs1992] (see below); David Lewis. 1992. Text Representation for Intelligent Text Retrieval: A Classification-Oriented View. In [Jacobs1992] (see below); Karen Sparck Jones. 1992. Assumptions and Issues in Text-Based Retrieval. In [Jacobs1992] (see below); Paul Jacobs (ed.) 1992. Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Lawrence Erlbaum Associates, Hillsdale, N.J. New Jersey; and Christos Faloutsos and Douglas Oard. 1996. A Survey of Document retrieval and Filtering Methods. Technical Report, Information Filtering Project, University of Maryland, College Park, Md.

Brief Summary Text (7):
The goal of an information extraction system, on the other hand, is to consult a corpus of documents, usually smaller than those involved in document retrieval tasks, and extract pre-specified items of information. Such a task might be defined, for instance, by specifying a template schema instances of which are to be filled automatically on the basis of a linguistic analysis of the texts in the corpus. For more information regarding a typical information extraction system, reference may be made to Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 Evaluation. In Proceedings of the 6th Message Understanding Conference, ARPA, Columbia, Md.

Brief Summary Text (8):
Work in the areas of document retrieval and information extraction has seen some success in their separate, distinct domains. However, successful integration of the two to create an information indexing and retrieval system has yet to be demonstrated. There is therefore a need for improving such document/content retrieval and information extraction technology.

Detailed Description Text (3):
The system 100 further includes an information extraction sub-system 104 which is capable of receiving the subset of documents from the document retrieval sub-system 102 for further processing. In particular, the information extraction sub-system 104 is adapted to consult the corpus of documents and extract prescribed items of information for refining the ranking process.

Detailed Description Text (5):
Thereafter, in operation 204, information is extracted from the ranked documents. The extracted information is then processed. See operation 206. The documents in the database are subsequently re-ranked based on results of the processing, as indicated in operation 208. The foregoing operations may optionally be executed by the information extraction sub-system 104 of FIG. 1.

Detailed Description Text (7):
In one embodiment, the extracted information may include segments, or phrases, in the documents. Further, the processing may include comparing the segments with a predetermined set of patterns such as word phrases that are known to be associated with a particular topic. The results of the processing may include a sum of a plurality of scores that are assigned based on the comparison. Additional information relating to such scoring process and the various rules

associated therewith will be set forth hereinafter in greater detail.

Detailed Description Text (58):
One embodiment of the information extraction sub-system may be based on a cascade of finite-state transducers that compute the transformation of text from sequences of characters to domain templates. Each transducer (or "phase") in the present embodiment takes the output of the previous phase and maps it into structures that constitute the input to the next phase, or in the case of the final phase, that contain the domain template information that is the output of the extraction process. A typical application might employ various sequences of phases, although the number of transducers in different applications may vary. Table 1 illustrates a non-exhaustive list of transducers.

Detailed Description Text (71):
Having extended the method of general rules and application-specific instances to the Parser and Combiner, the example focuses on the ability to write grammars for multiple topics. The topic associated with the present example includes: "Document will announce the appointment of a new CEO and/or the resignation of a CEO of a company." The foregoing topic was run in a database as an ad hoc query, producing a set of 1000 text documents it deemed most likely to be relevant, and ranking them in order from most likely relevant to least likely. Both the document set and the ordering served as inputs to information extraction sub-system.

Detailed Description Text (72):
In one embodiment, the information extraction sub-system may include FASTUS offered by SRI International. For more information on such system, reference may be made to: Doug Appelt, John Bear, Jerry Hobbs, David Israel, Megumi Kameyama, Andrew Kehler, Mark Stickel, and Mabry Tyson. 1995. SRI International's FASTUS System MUC-6 Test Results and Analysis. In Proceedings of the 6th Message Understanding Conference, ARPA, Columbia, Md.; and Doug Appelt, John Bear, Jerry Hobbs, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. 1996. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In Emmanuel Roche and Yves Schabes (eds.) Finite State Devices for Natural Language Processing. MIT Press, Cambridge, Mass., which are each incorporated herein by reference in their entirety. Further, it should be noted that the document retrieval sub-system may include SMART offered by GENERAL ELECTRIC.

Detailed Description Text (73):
Two different schemes were tried for using the information from the information extraction sub-system to reorder the input list. Both involved configuring the grammar to assign scores to phrases based on correlation of phrase type with relevance. In one scheme, scores were assigned to patterns manually, based on intuitions as to differential contributions to relevance judgments; in the second, a probabilistic model for the relevance of a document was inferred from a set of training data.

Detailed Description Text (74):
As a basis for the present experiment, 100 were picked articles from the middle of the ordered set that the document retrieval sub-system produced (in particular, articles ranked 401 through 500). The templates that the information extraction sub-system produced from those articles were examined to identify criteria for assigning a relevance rank to an article. The information extraction sub-system then assigned a numerical score from 0.1 to 1000 to the templates that it produced for a phrase. The manner in which this scoring is carried out is shown in Table 5.

Detailed Description Text (76):
For a second experiment, it was asked how a system for automatically identifying features concerning the output of the information extraction sub-system would compare with the results obtained by the manually tuned system. This was accomplished by first determining the relative strengths of the features concerning the output of the information extraction sub-system. A probabilistic model for the relevance of a document was inferred from a set of training data.

Detailed Description Text (77):
In yet another experiment, a ranked list of 2,000 documents for each query were used. Grammars

for 23 of the 47 topics were developed. For these 23 topics, the information extraction sub-system ran over the 2,000 articles, reordered them, and truncated the same to 1000. For the other 24 topics, the original ordering was truncated at 1000 documents. As in the first experiments set forth hereinabove, the reordering is achieved by having patterns, that is, instances (see example above), assign a score to the segment (phrase) of an article successfully matched against. An article's total score is the sum of the scores of all the patterns that matched against phrases in that article.

Detailed Description Text (79):
The scores were assigned with a threshold score in mind. An article had to contain at least one pattern that had a score of 1000 or above to be moved toward the beginning of the ordering; scores below 1000 had no effect on the order and were used solely for diagnostics. There was one exception to this general rule. In one topic, the following mini-experiment was used: phrases were assigned maximum scores of 250, so that at least four matching phrases were needed to move an article to the front of the ranking. This was intended to handle cases where one could find no especially reliable phrasal indicators of relevance. Another way to put this is that this method is a crude approximation to a statistical approach based on co-occurrence data. When writing the grammars, the present approach was to aim for high precision and to sacrifice recall when in a position to make a precision/recall tradeoff.

Detailed Description Text (83):
FIG. 5 illustrates a Content Exchange (CE) system 500 according to an embodiment of the present invention. The system includes all participating local systems with their associated Dialogue Managers, Extraction Engines and local Information Caches, (and any local structured databases the Dialogue Managers can query). At its center is a networked content directory, which contains representations of the information holdings of the sites in the Network and which is used to handle requests from local systems for content from the network.

Detailed Description Text (95):
Build a taxonomy keyed to specific sites and compare terms in the query to the taxonomy.

Detailed Description Text (201):
In another embodiment of the present invention, the analysis of the information about the content includes performing a linguistic analysis on the content of the at least one network data site for generating the index. Alternatively, the analysis of the information can include performing a statistical analysis of co-occurrence data, i.e., the matching of word and/or phrase patterns to a model word and/or phrase, to determine where to index a particular content item.

Detailed Description Paragraph Table (1):
TABLE 1 1. Tokenizer. This phase accepts a stream of characters as input, and transforms it into a sequence of tokens. 2. Multiword Analyzer. This phase is generated automatically by the lexicon to recognize token sequences (like "because of") that are combined to form single lexical items. 3. Name Recognizer. This phase recognizes word sequences that can be unambiguously identified as names from their internal structure (like "ABC Corp." and "John Smith"). 4. Parser. This phase constructs basic syntactic constituents of the language, consisting only of those that can be nearly unambiguously constructed from the input using finite-state rules (i.e., noun groups, verb groups, and particles). 5. Combiner. This phase produces larger constituents from the output of the parser when it can be done fairly reliably on the basis of local information. Examples are possessives, appositives, "of" prepositional phrases ("John Smith, 56, president of IBM's subsidiary"), coordination of same-type entities, and locative and temporal prepositional phrases. 6. Domain or Clause-Level Phase. The final phase recognizes the particular combinations of subjects, verbs, objects, prepositional phrases, and adjuncts that are necessary for correctly filling the templates for a given information extraction task.

Current US Original Classification (1):
707/101

Current US Cross Reference Classification (1):

707/102

Current US Cross Reference Classification (2):
707/3

Previous Doc Next Doc Go to Doc#

automated generation of user-specific recommendations. The system uses a statistical latent class model, also known as Probabilistic Latent Semantic Analysis, to integrate data including textual and other content descriptions of items to be searched, user profiles, demographic information, query logs of previous searches, and explicit user ratings of items. The disclosed system learns one or more statistical models based on available data. The learning may be reiterated once additional data is available. The statistical model, once learned, is utilized in various ways: to make predictions about item relevance and user preferences on un-rated items, to generate recommendation lists of items, to generate personalized search result lists, to disambiguate a users query, to refine a search, to compute similarities between items or users, and for data mining purposes such as identifying user communities.

13 Claims, 15 Drawing figures

☐ 3. Document ID: US 6651058 B1

L19: Entry 3 of 9 File: USPT Nov 18, 2003

US-PAT-NO: 6651058
DOCUMENT-IDENTIFIER: US 6651058 B1

TITLE: System and method of automatic discovery of terms in a document that are relevant to a given target topic

DATE-ISSUED: November 18, 2003

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Sundaresan; Neelakantan	San Jose	CA		
Yi; Jeonghee	San Jose	CA		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
International Business Machines Corporation	Armonk	NY			02

APPL-NO: 09/ 439758 [PALM]
DATE FILED: November 15, 1999

PARENT-CASE:
CROSS-REFERENCE TO RELATED APPLICATIONS This application relates to U.S. patent applications Ser. No. 09/440,625, is now U.S. Pat. No. 6,385,629 titled "System and Method for the Automatic Mining of Acronym-expansion Pairs Patterns and Formation Rules", Ser. No. 09/439,379, now pending titled "System and Method for the Automatic Mining of Patterns and Relations", Ser. No. 09/440,203, now pending titled "System and Method for the Automatic Construction of Generalization-Specialization Hierarchy of Terms", Ser. No. 09/440,602, titled "System and Method for the Automatic Recognition of Relevant Terms by Mining Link Annotations", and Ser. No. 09/440,626, is now pending titled "System and Method for the Automatic Mining of New Relationships", all of which are assigned to, and were filed by the same assignee as this application on even date herewith, and are incorporated herein by reference in their entirety.

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/6; 707/3, 707/5, 707/100, 707/102, 706/45, 706/50

US-CL-CURRENT: 707/6; 706/45, 706/50, 707/100, 707/102, 707/3, 707/5

FIELD-OF-SEARCH: 707/1-10, 707/100-104.1, 707/200-205, 707/501.1, 707/512-515, 709/200-235, 706/45-50

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
5642502	June 1997	Driscoll	707/200
5692107	November 1997	Simoudis et al.	706/12
5745360	April 1998	Leone et al.	364/140
5819260	October 1998	Lu et al.	707/3
5857179	January 1999	Vaithyanathan et al.	707/2
5933822	August 1999	Braden-Harder et al.	707/3
6006221	December 1999	Liddy et al.	704/2
6122647	September 2000	Horowitz et al.	707/3
6134532	October 2000	Lazarus et al.	705/1
6175829	January 2001	Li et al.	382/230
6182091	January 2001	Pitkow et al.	707/102
6185550	February 2001	Snow et al.	707/1
6377947	April 2002	Evans	707/5
6389436	May 2002	Chakrabarti et al.	707/3

OTHER PUBLICATIONS

R. Larson, "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace," the Proceedings of the 1966 American Society for Information Science Annual Meeting, also published as a technical report, School of Information Management and Systems, University of California, Berkeley, 1996, which is published on the Word Wide Web at URL: http://sherlock.sims.berkeley.edu/docs/asis96/asis96.html.
D. Gibson et al., "Inferring Web Communities from Link Topology," Proceedings of the 9.sup.th ACM. Conference on Hypertext and Hypermedia, Pittsburgh, PA, 1998.
D. Turnbull. "Bibliometrics and the World Wide Web," Technical Report University of Toronto, 1996.
K. McCain, "Mapping Authors in Intellectual Space: A technical Overview," Journal of the American Society for Information Science, 41(6):433-443, 1990.
S. Brin, "Extracting Patterns and Relations from the World Wide Web," WebDB, Valencia, Spain, 1998.
R. Agrawal et al., "Fast Algorithms for Mining Association Rules," Proc. of the 20th Int'l Conference on VLDB, Santiago, Chile, Sep. 1994.
R. Agrawal et al., Mining Association Rules Between Sets of Items in Large Databases, Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993.
S. Chakrabarti et al. "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," Proc. of The 8.sup.th International World Wide Web Conference, Toronto, Canada, May 1999.
B. Huberman et al., "Strong Regularities in Word Wide Web Surfing," Xerox Palo Alto Research Center.
A. Hutchunson, "Metrics on Terms and Clauses," Department of Computer Science, King's College London.
J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Proc. of 9th ACM-SIAM

Symposium on Discrete Algorithms, May 1997.
R. Srikant et al., "Mining Generalized Association Rules," Proceedings of the 21.sup.st VLDB Conference, Zurich, Swizerland, 1995.
W. Li et al., Facilitating comlex Web queries through visual user interfaces and query relaxation, published on the Word Wide Web at URL:
http://www.7scu.edu.au/programme/fullpapers/1936/com1936.htm as of Aug. 16, 1999.
G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules," pp. 229-248.
R. Miller et al., "SPHINX: A Framework for Creating Personal, Site-specific Web Crawlers," published on the Word Wide Web at URL:
http://www.7scu.edu.au/programme/fullpapers/1875/com1875.htm as of Aug. 16, 1999.
S. Soderland. Learning to Extract Text-based Information from the World Wide Web, American Association for Artificial Intelligence (www.aaai.org), pp. 251-254.
G. Plotkin. A Note Inductive Generalization, pp. 153-163.
R. Feldman et al., "Mining Associations in Text in the Presence of Background Knowledge," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Aug. 2-4, 1996, Portland, Oregon.
R. Kumar et al., "Trawling the Web for Emerging Cyber-Communities," published on the Word Wide Web at URL: http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html as of Nov. 13, 1999.
"Acronym Finder", published on the Word Wide Web at URL:http://acronymfinder.com/ as of Sep. 4, 1999.

ART-UNIT: 2177

PRIMARY-EXAMINER: Channavajjala; Srirama

ATTY-AGENT-FIRM: Kassatly; Samuel A.

ABSTRACT:

A computer program product is provided as an automatic mining system to discover terms that are relevant to a given target topic from a large databases of unstructured information such as the World Wide Web. The operation of the automatic mining system is performed in three stages: The first stage is carried out by a new terms discoverer for discovering the terms in a document, the second stage is carried out by a candidate terms discoverer for discovering potentially relevant terms, and the third stage is carried out by a relevant terms discoverer for refining or testing the discovered relevance to filter false relevance. The new terms discoverer includes a system for the automatic mining of patterns and relations, a system for the automatic mining of new relationships, and a system for selecting new terms from relations. In one embodiment, the system for the automatic mining of patterns and relations identifies a set of related terms on the WWW with a high degree of confidence, using a duality concept, and includes a terms database and two identifiers: a relation identifier and a pattern identifier. The system for the automatic mining of new relationships includes a database a knowledge module and a statistics module. The knowledge module includes a stemming unit, a synonym check unit, and a domain knowledge check unit. The candidate terms discoverer includes a metadata extractor, a document vector module, an association module, a filtering module, and a database. The relevant terms discoverer includes a stop word filter and a system for the automatic construction of generalization--specialization hierarchy of terms comprised of a terms database, an augmentation module, a generalization detection module, and a hierarchy database.

22 Claims, 9 Drawing figures

☐ 4. Document ID: US 6519602 B2

L19: Entry 4 of 9 File: USPT Feb 11, 2003

US-PAT-NO: 6519602
DOCUMENT-IDENTIFIER: US 6519602 B2

TITLE: System and method for the automatic construction of generalization-specialization hierarchy of terms from a database of terms and associated meanings

DATE-ISSUED: February 11, 2003

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Sundaresan; Neelakantan	San Jose	CA		
Yi; Jeonghee	San Jose	CA		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
International Business Machine Corporation	Armonk	NY			02

APPL-NO: 09/ 440203 [PALM]
DATE FILED: November 15, 1999

PARENT-CASE:
CROSS-REFERENCE TO RELATED APPLICATIONS This application relates to co-pending U.S. patent applications Ser. No. 09/440,625, is now U.S. Pat. No. 6,385,629 titled "System and Method for the Automatic Mining of Acronym-expansion Pairs Patterns and Formation Rules", Ser. No. 09/439,379, is now pending titled "System and Method for the Automatic Mining of Patterns and Relations", Ser. No. 09/440,602, is now pending titled "System and Method for the Automatic Recognition of Relevant Terms by Mining Link Annotations", Ser. No. 09/439,758, is now pending titled "System and Method for the Automatic Discovery of Relevant Terms from the World Wide Web", and Ser. No. 09/440,626, is now pending titled "System and Method for the Automatic Mining of New Relationships", all of which are assigned to, and were filed by the same assignee as this application on even date herewith, and are incorporated herein by reference in their entirety.

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/100; 707/1, 707/3, 707/6, 707/102, 707/513, 707/514
US-CL-CURRENT: 707/100; 707/1, 707/102, 707/3, 707/6, 715/513, 715/514

FIELD-OF-SEARCH: 707/1-8, 707/10, 707/100-104.1, 707/200-202, 707/501.1, 707/516, 707/907, 707/500.1, 707/512-514, 709/200-225

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
5745360	April 1998	Leone et al.	364/140
5748186	May 1998	Raman	707/500.1
5809499	September 1998	Wong et al.	178/18.01
5819260	October 1998	Lu et al.	707/3
5857179	January 1999	Vaithyanathan et al.	707/2
6122647	September 2000	Horowitz et al.	707/513

6128613	October 2000	Wong et al.	707/5
6128619	October 2000	Fogarasi et al.	707/10
6240407	May 2001	Chang et al.	707/1
6243700	June 2001	Zellweger	345/866
6279006	August 2001	Shigemi et al.	707/100

OTHER PUBLICATIONS

Hiroki Arimura et al AGeneralization of the Least general generalization, Machine Intelligence 13, pp. 59-85, 1994.*
Shan-Hwei Nienhuys-Cheng, Least generalizations and greatest specializations of sets of clauses, Journal of Artificial Intelligence Research, 1996, pp341-363.*
R. Larson, "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace," the Proceedings of the 1966 American Society for Information Science Annual Meeting, also published as a technical report, School of Information Management and Systems, University of California, Berkeley, 1996, which is published on the Word Wide Web at URL: http://sherlock.sims.berkeley.edu/docs/asis96/asis96.html.
D. Gibson et al., "Inferring Web Communities from Link Topology" Proceedings of the 9.sup.th ACM. Conference on Hypertext and Hypermedia, Pittsburgh, PA, 1998.
D. Turnbull. "Bibliometrics and the World Wide Web," Technical Report University of Toronto, 1996.
K. McCain, "Mapping Authors in Intellectual Space: A technical Overview," Journal of the American Society for Information Science, 41(6):433-443, 1990.
S. Brin, "Extracting Patterns and Relations from the World Wide Web," WebDB, Valencia, Spain, 1998.
R. Agrawal et al., "Fast Algorithms for Mining Association Rules," Proc. of the 20th Int'l Conference on VLBD, Santiago, Chile, Sep. 1994.
R. Agrawal et al., Mining Association Rules Between Sets of Items in Large Databases, Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993.
S. Chakrabarti et al., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," Proc. of The 8.sup.th International World Wide Web Conference, Toronto, Canada, May 1999.
B. Huberman et al., "Strong Regularities in World Wide Web Surfing," Xerox Palo Alto Research Center.
A. Hutchunson, "Metrics on Terms and Clauses," Department of Computer Science, King's College London.
J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Proc. of 9th ACM-SIAM Symposium on Discrete Algorithms, May 1997.
R. Srikant et al., "Mining Generalized Association Rules," Proceedings of the 21.sup.st VLDB Conference, Zurich, Switzerland, 1995.
W. Li et al., "Facilitating complex Web queries through visual user interfaces and query relaxation," published on the Word Wide Web at URL: http://www.7scu.edu.au/programme/fullpapers/1936/com1936.htm as of Aug. 16, 1999.
G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules," pp. 229-248.
R. Miller et al., "SPHINX: A Framework for Creating Personal, Site-specific Web Crawlers," published on the Word Wide Web at URL: http://www.7scu.edu.au/programme/fullpapers/1875/com1875.htm as of Aug. 16, 1999.
S. Soderland, "Learning to Extract Text-based Information from the World Wide Web," American Association for Artificial Intelligence (www.aaai.org), pp. 251-254.
G. Plotkin. "A Note Inductive Generalization," pp. 153-163.
R. Feldman et al., "Mining Associations in Text in the Presence of Background Knowledge," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Aug. 2-4, 1996, Portland, Oregon.
R. Kumar et al., "Trawling the Web for Emerging Cyber-Communities," published on the Word Wide Web at URL: http://www.8.org/w8-papers/4a-search-mining/trawling/trawling.html as of Nov. 13, 1999.
"Acronym Finder", published on the Word Wide Web at URL:http://acronymfinder.com/as of Sep. 4,

1999.

ART-UNIT: 2197

PRIMARY-EXAMINER: Channavajjala; Srirama

ATTY-AGENT-FIRM: Kassatly; Samuel A.

ABSTRACT:

A computer program product is provided as an automatic mining system to build a generalization hierarchy of terms from a database of terms and associated meanings, using for example the Least General Generalization (LGG) model. The automatic mining system is comprised of a terms database, an augmentation module, a generalization detection module, and a hierarchy database. The terms database stores the terms and their meanings, and the hierarchy database stores the generalization hierarchy which is defined by a set of edges and nodes. The augmentation module updates the terms using the LGG model. The generalization detection module maps the generalizations derived by the augmentation module, updates the edges, and derives a generalization hierarchy. In operation, the automatic mining system begins with no predefined taxonomy of the concept terms, and the LGG model derives a generalization hierarchy, modeled as a Directed Acyclic Graph from the terms.

9 Claims, 5 Drawing figures

☐ 5. Document ID: US 6442545 B1

L19: Entry 5 of 9 File: USPT Aug 27, 2002

US-PAT-NO: 6442545
DOCUMENT-IDENTIFIER: US 6442545 B1

TITLE: Term-level text with mining with taxonomies

DATE-ISSUED: August 27, 2002

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Feldman; Ronen	Petach Tikvah			IL
Aumann; Yehonatan	Jerusalem			IL
Schler; Jonathan	Petach Tikvah			IL
Landau; David	Rehovot			IL
Lipshtat; Orly	Jerusalem			IL
Ben-Yehuda; Yaron	Ramat Gan			IL

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Clearforest Ltd.	Petach Tikva			IL	03

APPL-NO: 09/ 323491 [PALM]
DATE FILED: June 1, 1999

PARENT-CASE:
MICROFICHE APPENDIX A computer printout is attached hereto as an appendix in microfiche form and is incorporated herein by reference. The printout comprises executable program files in hexadecimal format. This appendix includes 5 microfiches, containing a total of 424 frames.

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/6; 707/10
US-CL-CURRENT: 707/6; 707/10

FIELD-OF-SEARCH: 707/6, 707/10, 707/102, 707/104.1, 709/203

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
6233575	May 2001	Agrawal et al.	707/6

OTHER PUBLICATIONS

"Mining Text Using Keyword Distributions", by Ronen Feldman, et al., Proceedings of the 1995 Workshop on Knowledge Discovery in Databases.
"Finding Associations in Collections of Text", by Ronen Feldman, et al., Machine Learning and Data Mining: Methods and Applications, Edited by R.S. Michalski, et al., John Wiley & Sons, Ltd., 1997.
"Technology Text Mining, Turning Information Into Knowledge: A White Paper from IBM", Edited by Daniel Tkach, Feb. 17, 1998.
"Text Mining at the Term Level", by Feldman, et al., Proceedings of the 1998 Workshop on Knowledge Discovery in Databases, Aug. 1998.

ART-UNIT: 2175

PRIMARY-EXAMINER: Popovici; Dov

ASSISTANT-EXAMINER: Mofiz; Apu M

ATTY-AGENT-FIRM: Pennie & Edmonds LLP

ABSTRACT:

A method for mining in a database including documents, the documents including text. The method includes providing a taxonomy of taxonomy terms, and mining the documents responsive to the taxonomy to discover a relationship between a set of one or more selected words and at least one of the taxonomy terms. The method also includes analyzing occurrences of the relationship over a plurality of the documents to extract information relating to the at least one taxonomy term.

59 Claims, 8 Drawing figures

☐ 6. Document ID: US 6430557 B1

L19: Entry 6 of 9 File: USPT Aug 6, 2002

US-PAT-NO: 6430557
DOCUMENT-IDENTIFIER: US 6430557 B1

TITLE: Identifying a group of words using modified query words obtained from successive suffix relationships

DATE-ISSUED: August 6, 2002

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Gaussier; Eric	Grenoble			FR
Grefenstette; Gregory	Gieres			FR
Chanod; Jean-Pierre	Grenoble			FR

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Xerox Corporation	Stamford	CT			02

APPL-NO: 09/ 212662 [PALM]
DATE FILED: December 16, 1998

INT-CL: [07] G06 F 17/30, G06 F 17/27, G06 F 17/21

US-CL-ISSUED: 707/5; 707/6, 704/9, 704/10
US-CL-CURRENT: 707/5; 704/10, 704/9, 707/6

FIELD-OF-SEARCH: 707/1-7, 704/1-10, 704/205

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
4799188	January 1989	Yoshimura	364/900
4864501	September 1989	Kucera et al.	364/419
5488725	January 1996	Turtle et al.	707/5
5551049	August 1996	Kaplan et al.	395/800
5594641	January 1997	Kaplan et al.	395/601
5625554	April 1997	Cutting et al.	707/100
5696962	December 1997	Kupiec	707/4
5940624	August 1999	Kadashevich et al.	704/9
5963940	October 1999	Liddy et al.	707/5
6012053	January 2000	Pant et al.	707/3
6081774	June 2000	de Hita et al.	704/9
6092065	July 2000	Floratos et al.	707/6
6101492	August 2000	Jacquemin et al.	707/3
6105023	August 2000	Callan	707/5
6308149	October 2001	Gaussier et al.	704/9

OTHER PUBLICATIONS

Adamson, George W. et al. "The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles," Inform. Stor. Retr. vol. 10, 1974, pp. 253-260.
Chanod, Jean-Pierre et al. "Taging French-Comparing a Statistical and a Constraint-Based Method," Rank Xerox Research Centre, pp. 149-156.
Dawson, J.L. "Suffix Remvoal and Word Conflation," ALLC Bulletin, 1974, pp. 33-46.
Frakes, W.B. "Stemming Algorithms," Information Retrieval Data Structures & Algorithms, Prentice Hall, New Jersey, 1992, pp. 131-160.
Frakes, W.B. "Term Conflation for Information Retrieval," Research and Development in Information Retrieval Proceedings of the Third Joint BCS and ACM Symposium, King's College, Cambridge, Jul. 2-6, 1984, pp. 383-389.
Hafer, Margaret A. et al. "Word Segmentation by Letter Successor Varieties," Inform. Stor, Retr. vol. 10, pp. 371-385.
Harman, Donna "How Effective Is Suffixing?" Journal of the American Society for Information Science, Jan. 1991, vol. 42, No. 1, pp. 7-15.
Hull, David A. "Stemming Algorithms: A Case Study for Detailed Evaluation," Journal of the American Society for Information Science, Jan. 1996, vol. 47, No. 1, pp. 70-84.
Jacquemin, Christian "Guessing Morphology from Terms and Corpora," Proceedings of the 20.sup.th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, Jul. 27-31, 1997, pp. 156-165.
Karttunen, Lauri et al. "A Compiler for Two-Level Phonological Rules," Jun. 25, 1987.
Karttunen, Lauri "Constructing Lexical Transducers," Proceedings of the 15.sup.th International Conference on Computational Linguistics, Kyoto, Japan, Aug. 5-9, 1994, pp. 406-411.
Karttunen, Lauri "The Replace Operator," Proceedings of 33.sup.rd Annual Meeting of the Association for Computational Linguistics, ACL-94, Boston, Massachusetts, 1995, pp. 16-23.
Kaplan, Ronald M. et al. "Regular Models of Phonological Rule Systems,"Computational Linguistics vol. 20, No. 3, 1994, pp. 331-380.
Lennon, Martin et al. "An Evaluation of Some Conflation Algorithms for Information Retrieval," Journal of Information Science 3, 1981, pp. 177-183.
Lovins, Julie Beth "Development of a Stemming Algorithm*,"Mechanical Translation and Computational Linguistics, vol. 11, 1968, pp. 22-31.
Paice, Chris D. "Another Stemmer," SIGIR Forum, Fall 1990, vol. 24, No. 3, pp. 56-61.
Paice, Chris D. "Method for Evaluation of Stemming Algorithms Based on Error Counting," Journal of the American Society for Information Science, vol. 47, No. 8, Aug. 1996, pp. 632-649.
Porter, M.F. "An Algorithm for Suffix Stripping," Program Automated Library and Information Systems, vol. 14, No. 3, Jul. 1980, pp. 130-137.
Rasmussen, Edie "Clustering Algorithms," pp. 419-436.
Roche, Emmanuel et al. "Deterministic Part-of-Speech Tagging with Finite-State Transducers," Computational Linguistics, vol. 21, No. 2, 1995, pp. 227-253.
Romesburg, H. Charles Cluster Analysis for Researchers, 1984, pp. 9-15.
Salton, Gerard et al. Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983, pp. 130-136.
"Xerox Linguistic Database Reference English Version 1.1.4," Xerox Corporation, 1994.

ART-UNIT: 2161

PRIMARY-EXAMINER: Trammell; James P.

ASSISTANT-EXAMINER: Wang; Mary

ATTY-AGENT-FIRM: Oliff & Berridge, PLC

ABSTRACT:

A query word is used to identify one of a number of word groups, by first determining whether the query word is in any of the word groups. If not, attempts to modify the query word are made in accordance with successive suffix relationships in a sequence until a modified query word is

obtained that is in one of the word groups. The sequence of suffix relationships, which can be pairwise relationships, can be defined by a list ordered according to the frequencies of occurrence of the suffix relationships in a natural language. If a modified query word is obtained that is in one of the word groups, information identifying the word group can be provided, such as a representative of the group or a list of words in the group.

20 Claims, 7 Drawing figures

☐ 7. Document ID: US 6154213 A

L19: Entry 7 of 9 File: USPT Nov 28, 2000

US-PAT-NO: 6154213
DOCUMENT-IDENTIFIER: US 6154213 A

TITLE: Immersive movement-based interaction with large complex information structures

DATE-ISSUED: November 28, 2000

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Rennison; Earl F.	San Francisco	CA	94107	
Strausfeld; Lisa S.	San Francisco	CA	94109	
Horowitz; Damon M.	San Francisco	CA	94117	

APPL-NO: 09/ 087259 [PALM]
DATE FILED: May 29, 1998

PARENT-CASE:
RELATED APPLICATION This application is a continuation of Serial No. 60/048,150, entitled "Immersive Movement-Based Interaction with Large Complex Information Structures" filed on May 30, 1997 pending, which is incorporated in its entirety by reference herein, and which is assigned to a common assignee as the present application.

INT-CL: [07] G06 F 3/14

US-CL-ISSUED: 345/356; 345/357, 345/334, 345/349, 345/428, 345/333, 707/103, 707/104, 707/501
US-CL-CURRENT: 715/854; 345/428, 707/103R, 707/104.1

FIELD-OF-SEARCH: 345/356, 345/353, 345/357, 345/348, 345/349, 345/333, 345/334, 345/428, 707/103, 707/104, 707/501, 707/514

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
5008853	April 1991	Bly et al.	345/331
5062060	October 1991	Kolnick	345/339
5241671	August 1993	Reed et al.	707/104

5481666	January 1996	Nguyen et al.	345/357
5537526	July 1996	Anderson et al.	707/515
5544302	August 1996	Nguyen	345/348
5550563	August 1996	Matheny et al.	345/348
5557722	September 1996	Derose et al.	345/357
5584035	December 1996	Duggan et al.	345/339
5623589	April 1997	Needham et al.	707/501
5675752	October 1997	Scott et al.	345/352
5721851	February 1998	Cline et al.	345/357
5832494	November 1998	Egger et al.	707/104
5877766	March 1999	Bates et al.	345/357
5978811	November 1999	Smiley	707/104

ART-UNIT: 273

PRIMARY-EXAMINER: Bayerl; Raymond J.

ASSISTANT-EXAMINER: Nguyen; Thomas T.

ATTY-AGENT-FIRM: Fenwick & West LLP

ABSTRACT:

An intuitive, immersive, movement-based interface and system provides for navigating through large collections of multidimensional information. The interface allows users to navigate through large document collections by maintaining a constant density of visual information presented on a display device to the user at any given moment of time. The document collection is organized in an immersive information space, containing various levels of topics and related documents. At each level within the immersive information space contextual information is presented to the user. The contextual information consists of a semantic scale and a pathway to the information they are viewing. An information structure represents the immersive information space of documents. The information structure consists of a collection of documents, and a graph of topics that describe the relationships between the documents. The graph of topics consists of topic nodes that each contain 1) a set of documents that are about that topic, and 2) a set of links to other topics in the structure. The links represent relationships between topics, and indirectly, relationships between the documents. An information structure that represents the collection of documents is used to guide the user to documents of interest and to show relationships between documents. A presentation and interaction model allows navigation through the information structure. The model includes a camera representing a user's focus of attention, and a set of reactable graphical objects representing nodes in the information structure. The interaction model continuously monitors the movement of the camera in relation to the graphical objects and updates the display of the information space.

11 Claims, 10 Drawing figures

☐ 8. Document ID: US 5873056 A

L19: Entry 8 of 9 File: USPT Feb 16, 1999

US-PAT-NO: 5873056
DOCUMENT-IDENTIFIER: US 5873056 A

TITLE: Natural language processing system for semantic vector representation which accounts for lexical ambiguity

DATE-ISSUED: February 16, 1999

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Liddy; Elizabeth D.	Syracuse	NY		
Paik; Woojin	Syracuse	NY		
Yu; Edmund Szu-li	Syracuse	NY		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
The Syracuse University	Syracuse	NY			02

APPL-NO: 08/ 135815 [PALM]
DATE FILED: October 12, 1993

INT-CL: [06] G06 F 17/30, G06 F 17/20, G06 F 17/22

US-CL-ISSUED: 704/9; 707/1, 707/3, 707/101, 707/532
US-CL-CURRENT: 704/9; 707/1, 707/101, 707/3, 715/532

FIELD-OF-SEARCH: 395/600, 395/12, 395/63, 395/934, 364/419.01, 364/419.08, 364/419.13, 707/1, 707/104, 707/530, 707/532, 707/3, 707/101, 704/9, 704/10

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
4358824	November 1982	Glickman et al.	364/419.19
4495566	January 1985	Dickinson et al.	395/600
4580218	April 1986	Raye	364/419.13
4803642	February 1989	Muranaga	395/62
4823306	April 1989	Barbic et al.	395/600
4839853	June 1989	Deerwester et al.	395/600
4849898	July 1989	Adi	364/419.1
4868733	September 1989	Fujisawa et al.	395/600
4972349	November 1990	Kleinberger	395/144
4994967	February 1991	Asakawa	364/419.08
5020019	May 1991	Ogawa	395/600
5056021	October 1991	Ausborn	364/419.08
5099426	March 1992	Carlgren et al.	364/419.13
5122951	June 1992	Kamiya	364/419.13
5128865	July 1992	Sadler	364/419.02
5140692	August 1992	Morita	395/600
5146405	September 1992	Church	364/419.08
5151857	September 1992	Matsui	364/419.13
5162992	November 1992	Williams	364/419.1

5168565	December 1992	Morita	395/600
5197005	March 1993	Shwartz et al.	364/419.13
5237503	August 1993	Bedecarrax et al.	364/419.18
5285386	February 1994	Kuo	364/419.02
5297039	March 1994	Kanaegami et al.	364/419.13
5325298	June 1994	Gallant	364/419.19
5331556	July 1994	Black, Jr. et al.	364/419.08
5371807	December 1994	Register et al.	382/14
5418951	May 1995	Damashek	707/5
5541836	July 1996	Church et al.	704/7
5619709	April 1997	Caid et al.	707/532
5675819	October 1997	Schueteze	704/10
5694592	December 1997	Driscoll	395/603

OTHER PUBLICATIONS

Meteer et al, "POST: Using Probabilities in Language Processing," Proc. 12th Intl. Conf. on A.I. vol. 12, Aug. 1991, pp. 960-964.
Liddy et al, Proc. Workshop on Natural Language Learning, IJCAI, Sydney, Australia, 1991, pp. 50-57, entitled "An Intelligent Seimantic Relation Assignor: Preliminary Work.".
Stephen I. Gallant, "A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Network," Neural Computation 3, pp. 293-309, Massachusetts Institute of Technology, 1991.
Yorick Wilks et al., "Providing Machine Tractable Dictionary Tools," Machine Translation, pp. 98-154, Jun. 1990.
Gerard Salton et al., Introduction to Modern Information Retreival, Mc-Graw-Hill Book Company, pp. 118-155, Apr. 1983.
Ellen M. Voorhees et al., "Vector Expansion in a Large Collection," Siemans Coporate Research, Inc., Princeton, New Jersey, Unknown.
Scott Deerwester et al., "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, 41(6), pp. 391-407, 1990.
Hinrich Schutze, "Dimensions of Meaning," Proceedings Supercomputer '92, IEEE, pp. 787-796, Nov. 1992.
Gregory Grefenstette, "Use of Syntactic Context to Produce Term Association Lists for Text Retrieval," 18th Ann Int'l SIGIR '92, ACM, pp. 89-97, Jun. 1992.
Susan T. Dumais, "LSI meets TREC: A Status Report," NIST Special Publication 500-207, The First Text REtrieval Conference (TREC-1), pp. 137-152, Mar. 1993.
Elizabeth D. Liddy et al., "Statistically Guided Word Sense Disambiguation," Proceedings of the AAAI Fall 1992 Symposium on Probalistic Approach to Natural Language Processing, pp. 98-107, Oct. 1992.
Elizabeth D. Liddy et al., "Use of Subject Field Codes from a Machine-Readable Dictionary for Automatic Classification of Documents," Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop, Pittsburgh, PA, pp. 83-100, Oct. 1992.
Elizabeth D. Liddy et al., "DR-Link's Linguistic Conceptual Approach to Document Detection," Proceedings of TExt Retrieval Conference (TREC), 13 pages, Nov. 1992.
Elizabeth D. Liddy et al., "An Overview of DR-Link and its Approach to Document Filtering," Proceedings of the Human Language and Technology Workshop, 5 pages, Mar. 1993.

ART-UNIT: 276

PRIMARY-EXAMINER: Kulik; Paul V.

ATTY-AGENT-FIRM: Lukacher; K. J. Lukacher; M.

ABSTRACT:

A natural language processing system uses unformatted naturally occurring text and generates a subject vector representation of the text, which may be an entire document or a part thereof such as its title, a paragraph, clause, or a sentence therein. The subject codes which are used are obtained from a lexical database and the subject code(s) for each word in the text is looked up and assigned from the database. The database may be a dictionary or other word resource which has a semantic classification scheme as designators of subject domains. Various meanings or senses of a word may have assigned thereto multiple, different subject codes and psycholinguistically justified sense meaning disambiguation is used to select the most appropriate subject field code. Preferably, an ordered set of sentence level heuristics is used which is based on the statistical probability or likelihood of one of the plurality of codes being the most appropriate one of the plurality. The subject codes produce a weighted, fixed-length vector (regardless of the length of the document) which represents the semantic content thereof and may be used for various purposes such as information retrieval, categorization of texts, machine translation, document detection, question answering, and generally for extracting knowledge from the document. The system has particular utility in classifying documents by their general subject matter and retrieving documents relevant to a query.

46 Claims, 11 Drawing figures

☐ 9. Document ID: US 5809499 A

L19: Entry 9 of 9 File: USPT Sep 15, 1998

US-PAT-NO: 5809499
DOCUMENT-IDENTIFIER: US 5809499 A

TITLE: Computational method for discovering patterns in data sets

DATE-ISSUED: September 15, 1998

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Wong; Andrew K. C.	Waterloo			CA
Chau; Tom Tak Kin	Toronto			CA
Wang; Yang	Waterloo			CA

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
Pattern Discovery Software Systems, Ltd.	Waterloo			CA	03

APPL-NO: 08/ 733576 [PALM]
DATE FILED: October 18, 1996

INT-CL: [06] G06 F 17/30

US-CL-ISSUED: 707/6; 707/3, 395/12, 395/705, 704/1, 704/200, 345/156, 345/326, 345/339, 178/18
US-CL-CURRENT: 707/6; 178/18.01, 345/156, 704/1, 704/200, 706/11, 706/45, 707/104.1, 707/3, 715/700

FIELD-OF-SEARCH: 707/6, 707/3, 395/207, 395/209, 395/676, 395/751, 395/761, 395/705, 395/12, 704/1, 704/200, 345/156, 345/326, 345/339, 178/18

PRIOR-ART-DISCLOSED:

OTHER PUBLICATIONS

Smyth et al., "An information theoretic approach to rule induction from database," IEEE, pp. 301-316, Aug. 1992.
Langley et al., "Approaches to machine learning," IEEE, pp. 306-316, Sep. 1984.
Michalski et al., "Automated construction of classifications: conceptual clustering versus numerical taxonomy," IEEE, pp. 396-410, Jul. 1983.
Michalski et al., "Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology," IEEE, pp. 63-87, Jun. 1979.
Chan et al., "APACS: A system for Automatited pattern analysis and classification," computational Intelligence, vol. 6, No. 3, pp. 119-131, May, 1989.
D.H. Fischer, "Conceptual clustering, learning from examples, and inference", proceedings of the 4th Intenational workshop on machine learning, pp. 38-49, Jan. 1987.
Langey et al., "Conceptual clustering as discrimination learning", Proceedings of the Fifth Biennial Conference of the Canadian Society for computational Studies of Intelligence, pp. 95-98, Jan. 1984.
Fisher et al., "An empirical comparison of ID3 and Back-propagation", Proceedings of the 11th International joint conference on artificial Intelligence, vol. 1, pp. 788-793, Aug. 1989.
S.J. Harberman, "The analysis of residuals in cross-classified tables", Biometrics, vol.29, No. 1-4, pp. 205-220, Jan. 1973.
Shelby J. Haberman "The Analysis of reiduals in cross-classified tables", Biometrics, vol. 29, No. 1-4, pp. 205-220, Mar. 1973.
Douglas H. Fisher, "Knowledge acquisition via Incremental conceptual clustering", Machine Learning, vol.2, No.2, pp. 139-172, May 1987.
J.R. Quinlan, "Induction of Decision Trees", Machine Learning, V.1, No.1, pp. 81-106, Jun. 1986.
D.H. Fisher, "A Hierarchical conceptual Clustering Algorithm", Technical report, Dept of information and computer science, University of California, Irvine, Mar. 1984.
C. Berge, "Hypergraphes, combinatoires des ensembles finis", Annales de L'IHP--Analyse Non Lineaire, Jan. 1987.
Holsheimer et al., "Data mining the search for knowledge in databases", Computer Science/Department of Algorithmics and Architecture, Jan. 1994.

ART-UNIT: 271

PRIMARY-EXAMINER: Black; Thomas G.

ASSISTANT-EXAMINER: Corrielus; Jean M.

ATTY-AGENT-FIRM: Schumacher; Lynn C. Hill & Schumacher Dowell & Dowell, P.C.

ABSTRACT:

Automatic discovery of qualitative and quantitative patterns inherent in data sets is accomplished by use of a unified framework which employs adjusted residual analysis in statistics to test the significance of the pattern candidates generated from data sets. This framework consists of a search engine for different order patterns, a mechanism to avoid exhaustive search by eliminating impossible pattern candidates, an attributed hypergraph (AHG) based knowledge representation language and an inference engine which measures the weight of evidence of each pattern for classification and prediction. If a pattern candidate passes the statistical significance test of adjusted residual, it is regarded as a pattern and represented by an attributed hyperedge in AHG. In the task of classification and/or prediction, the weights of evidence are calculated and compared to draw the conclusion.

4 Claims, 1 Drawing figures

Full | Title | Citation | Front | Review | Classification | Date | Reference | Claims | KWIC | Draw. Desc | Image

Clear | Generate Collection | Print | Fwd Refs | Bkwd Refs | Generate OACS

Term	Documents
TAXONOMY	2190
TAXONOMIES	131
TAXONOMYS	0
(18 AND TAXONOMY).USPT.	9
(L18 AND TAXONOMY).USPT.	9

Display Format: FRO Change Format

Previous Page Next Page Go to Doc#